

1 **Tropospheric ozonesonde profiles at long-term U.S. monitoring sites: 1. A climatology**  
2 **based on self-organizing maps**

3 Ryan M. Stauffer<sup>1,2</sup>, Anne M. Thompson<sup>2,3</sup>, George S. Young<sup>2</sup>

4 28 December 2015

5 <sup>1</sup>Earth System Science Interdisciplinary Center (ESSIC), University of Maryland – College Park,  
6 College Park, Maryland, USA

7 <sup>2</sup>Department of Meteorology, The Pennsylvania State University, University Park, Pennsylvania,  
8 USA

9 <sup>3</sup>Earth Sciences Division, NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

10

11 Correspondence to: R. M. Stauffer ([rms5539@psu.edu](mailto:rms5539@psu.edu))

12

13 **Keywords/Index Terms**

14

15 Tropospheric Ozone – Ozonesondes – Climatology – CONUS Ozone – Self-organizing maps –  
16 STE

17

18 **Key Points**

- 19 • O<sub>3</sub> profile clusters correspond to large-scale meteorological conditions  
20 • Three clusters contain O<sub>3</sub> >100 ppbv above climatology near the tropopause  
21 • Clustering captures O<sub>3</sub> profile variability better than climatological means

22

23 **Abstract**

24

Sonde-based climatologies of tropospheric ozone ( $O_3$ ) are vital for developing satellite retrieval algorithms and evaluating chemical transport model output. Typical  $O_3$  climatologies average measurements by latitude or region, and season. Recent analysis using self-organizing maps (SOM) to cluster ozonesondes from two tropical sites found clusters of  $O_3$  mixing ratio profiles are an excellent way to capture  $O_3$  variability and link meteorological influences to  $O_3$  profiles. Clusters correspond to distinct meteorological conditions, e.g. convection, subsidence, cloud cover, and transported pollution. Here, the SOM technique is extended to four long-term U.S. sites (Boulder, CO; Huntsville, AL; Trinidad Head, CA; Wallops Island, VA) with 4530 total profiles. Sensitivity tests on k-means algorithm and SOM justify use of 3x3 SOM (nine clusters). At each site, SOM clusters together  $O_3$  profiles with similar tropopause height, 500 hPa height/temperature, and amount of tropospheric and total column  $O_3$ . Cluster means are compared to monthly  $O_3$  climatologies. For all four sites, near-tropopause  $O_3$  is double (over +100 parts per billion by volume; ppbv) the monthly climatological  $O_3$  mixing ratio in three clusters that contain 13 – 16% of profiles, mostly in winter and spring. Large mid-tropospheric deviations from monthly means (-6 ppbv, +7 – 10 ppbv  $O_3$  at 6 km) are found in two of the most populated clusters (combined 36 – 39% of profiles). These two clusters contain distinctly polluted (summer) and clean  $O_3$  (fall-winter, high tropopause) profiles, respectively. As for tropical profiles previously analyzed with SOM,  $O_3$  averages are often poor representations of U.S.  $O_3$  profile statistics.

## **1. Introduction**

### **1.1. Ozone Climatologies**

48        Since the 1960s, the global ozonesonde network has provided a comprehensive O<sub>3</sub> dataset  
49 of increasing spatial coverage and density, as well as recent quantification of short timescale  
50 processes such as pollution transport (e.g. Cooper et al., 2011) through campaign-based networks  
51 (Thompson et al., 2011). Campaign sonde networks also capture the evolution of stratosphere-to-  
52 troposphere exchange (STE; Holton et al., 1995; Lin et al., 2012) events. STE greatly affects the  
53 O<sub>3</sub> profile shape on short timescales with pronounced O<sub>3</sub> and potential vorticity correlations in the  
54 upper troposphere (Danielsen, 1968; Rao et al., 2003). Ozonesondes provide the highest vertical  
55 resolution measurements of O<sub>3</sub> available from the surface to above 30 km, at accuracies as high as  
56 ±5% (Komhyr et al., 1995). For these reasons, ozonesondes are the preferred reference  
57 measurements with which to compare chemical model output and satellite O<sub>3</sub> profile and column  
58 retrievals.

59        There have been efforts to establish global O<sub>3</sub> climatologies for comparison with model  
60 output and satellite measurements. These studies have relied heavily on the global ozonesonde  
61 network, using climatology as a baseline for trends (Logan, 1985, 1994; Logan et al. 1999, 2012;  
62 Oltmans et al., 2006, 2013), O<sub>3</sub> distribution in latitudinal bands (Stevenson et al., 2006), and  
63 climatology for specific regions (Newchurch et al., 2003; Tilmes et al., 2012) including the tropics  
64 (Thompson et al. 2003a,b; Thompson et al., 2012). Climatologies from ozonesondes and satellite  
65 retrievals in the stratosphere have also been developed to increase accuracy of total column O<sub>3</sub>  
66 integration from ozonesonde profiles (McPeters et al., 1997; McPeters and Labow, 2012). Model  
67 and satellite performance is judged primarily on replication of seasonal variability at one location  
68 or region, particularly in the upper troposphere/lower stratosphere (UTLS; e.g. Considine et al.,  
69 2008).

Recently, Tilmes et al. (2012) assembled a global O<sub>3</sub> climatology from 42 ozonesonde sites from 1980 – 1994 and 1995 – 2011. Their analysis separated the stations into 12 regions that exhibited similar O<sub>3</sub> probability density functions (PDFs). They demonstrated an application of the climatology via the improvements in CAM-Chem (Lamarque et al., 2012) model simulations that used derived stratospheric O<sub>3</sub>, as opposed to a monthly and latitudinally invariant stratospheric O<sub>3</sub> climatology. There are many processes however, such as synoptic-scale wave dynamics, which can cause significant deviations in the profile from a typical O<sub>3</sub> climatology. Thus, an investigation into these processes is performed using a technique that tends to classify O<sub>3</sub> profiles according to meteorological conditions and other influences on tropospheric O<sub>3</sub> profile shape.

## **1.2. Ozone Profile Clustering**

Two studies in particular set a precedent for clustering ozonesonde profiles. Diab et al. (2004) classified over 100 ozonesonde profiles launched from late 1998 – 2002 from a subtropical Southern Hemisphere Additional Ozonesondes (SHADOZ; Thompson et al., 2003a) site, Irene, South Africa. Their analysis yielded 6 clusters including distinct “background” and “polluted” clusters, containing well below, and well above average tropospheric O<sub>3</sub> mixing ratios. Diab et al. (2004) also found a cluster containing 48% of all profiles, which could not be ascribed to a particular meteorological regime, season, or source region. They labeled this cluster as representative of “typical” Irene O<sub>3</sub>, arguing that the representative cluster is more informative and descriptive of Irene O<sub>3</sub> than a mean profile because it is not influenced by extreme values and is not necessarily confined to a particular season. Their clustering analysis also allowed identification of STE/low tropopause height O<sub>3</sub> profiles, the majority of which occurred during the

Southern Hemisphere winter, when influences more characteristic of the mid-latitudes, namely the subtropical jet, frequently affect Irene.

Jensen et al. (2012) performed a cluster analysis on over 900 tropical ozonesonde profiles. They employed self-organizing maps (SOM; Kohonen, 1995), which have been used as a clustering algorithm across many disciplines, including several recent meteorology and climate studies (Hewitson and Crane, 2002; Hong et al., 2004; Liu et al., 2006; Nowotarski and Jensen, 2013). Jensen et al. (2012) used SOM to describe influences dictating O<sub>3</sub> variability at two SHADOZ stations, Natal, Brazil, and Ascension Island, from 1998 – 2009. Their four-cluster results were similar to those found in Diab et al. (2004), and were dominated by the seasonal influences of biomass burning and convection. Clusters representing a background state, a polluted state, and a mean state cluster with a plurality of profiles were found at both locations. The polluted clusters corresponded to the African biomass burning season in the Southern Hemisphere spring, leading to sharp O<sub>3</sub> gradients above the boundary layer and large mid-tropospheric O<sub>3</sub> amounts. The clean clusters contained launches primarily in the convective season, during which near-surface, low O<sub>3</sub> tropical air is lifted into the free troposphere.

Because of such source and synoptic effects that govern O<sub>3</sub> profile evolution throughout the year, clusters of O<sub>3</sub> profiles may be a better way to present a site's O<sub>3</sub> profile variability than monthly or seasonal averages. Thus, we are motivated to extend these techniques to data from several long-term mid-latitude ozonesonde sites. Following the approach of Jensen et al. (2012), SOM is applied to Contiguous United States (CONUS) tropospheric ozonesonde profiles. CONUS represents a somewhat confined, but varied geographic area, with thousands of high quality O<sub>3</sub> profiles from decades-long records available. The extremes of short-term vertical variability of

O<sub>3</sub> in the mid-latitudes are much greater than in the tropics, presenting a new challenge in interpreting the O<sub>3</sub> clustering statistics.

Our goals for this study are to: 1) Cluster tropospheric O<sub>3</sub> mixing ratio profiles at four CONUS sites using SOM. SOM is also evaluated against the similar k-means clustering algorithm to determine which method to apply in this paper. Sensitivity tests comparing the two methods and supporting the decision to use SOM are presented in the Appendix, as is as a technical discussion of both methods. 2) Provide meteorological and geophysical interpretations of SOM clusters and organization. Although chemical processes probably play a role in SOM classification, a scarcity of co-located trace gas data precludes characterization of chemistry's added influence. 3) Evaluate O<sub>3</sub> variability at each site by assessing the representativeness of a monthly O<sub>3</sub> profile climatology, focusing on deviations from the monthly climatology, mid-tropospheric O<sub>3</sub>, and the near-tropopause region.

## **2. Sonde Measurements and Analysis Techniques**

There are four ozonesonde sites with records of more than 15 years in CONUS: Boulder, CO; Huntsville, AL; Trinidad Head, CA; and Wallops Island, VA (Table 1). Newchurch et al. (2003) examined these four sites and compiled the first CONUS ozonesonde climatology using data from April 1995 to March 2002. Our analysis extends their dataset back in time to include the beginning of the Boulder and Wallops Island records, and adds a decade of observations beyond 2002 to each of the four sites. In this study, variability of O<sub>3</sub> is described without the constraints of monthly averages; rather they are used as context in this case. This method filters

background or polluted O<sub>3</sub> cases, and variations in tropopause height that are otherwise diluted by averaging.

The CONUS locations in this study span about 6° of latitude, and each is surrounded by unique terrain and experiences different regional influences. Boulder is just downwind of the Rocky Mountains and part of the Denver metro area. Huntsville, the southernmost site, is located in the southeast U.S. and exhibits more subtropical characteristics compared to the other sites (Newchurch et al., 2003; Tilmes et al., 2012). Trinidad Head is located on the coast of northern CA, is influenced by marine air masses from the Pacific, and of the four CONUS sites, is impacted most frequently by STE (Newchurch et al., 2003). Wallops Island is located on the Atlantic coast of the Delmarva Peninsula in southeast VA, often downwind of large emissions sources in the Ohio River Valley and the Baltimore/Washington D.C. region. For reference, Tilmes et al. (2012) combined Huntsville and Wallops Island into an “Eastern U.S.” region, while Boulder and Trinidad Head each remained isolated and unassigned to a regional grouping of sonde stations. Much like Tilmes et al. (2012), O<sub>3</sub> profile shapes and distributions are used to characterize ozonesonde sites. However, instead of developing a new O<sub>3</sub> climatology from SOM, the tendency of an O<sub>3</sub> climatology to describe clusters of O<sub>3</sub> profiles is evaluated. This is accomplished through use of a stricter, monthly O<sub>3</sub> climatology as opposed to a seasonal one as in Tilmes et al. (2012).

## **2.1. Data**

Ozonesonde data from the four CONUS locations in Table 1 were accessed through either the World Ozone and Ultraviolet Radiation Data Centre (WOUDC; <ftp://ftp.tor.ec.gc.ca/pub/woudc/>; Wallops Island, VA; portions of Boulder, CO) or NOAA Earth

160 System Research Laboratory Global Monitoring Division (ESRL GMD;  
161 <ftp://ftp.cmdl.noaa.gov/data/ozwv/Ozonesonde/>; portions of Boulder, CO; Huntsville, AL;  
162 Trinidad Head, CA). Ozonesondes are launched approximately weekly with each month of the  
163 year well represented at all sites (Figure 1). There are occasional increases in frequency for  
164 measurement campaigns, resulting in a total sample size of 4530 profiles.

165       The ozonesondes in this study use the electrochemical cell (ECC) instrument and  
166 processing technique described in Komhyr (1969). Typical uncertainties of ozonesonde  
167 measurements range from -7 to +17% in the troposphere, to  $\pm 5\%$  in the stratosphere (Komhyr et  
168 al., 1995). All profiles include measurements of pressure, temperature, and O<sub>3</sub> partial pressure. If  
169 not included with the standard measurements provided for each profile, variables such as O<sub>3</sub>  
170 mixing ratio, geopotential altitude, and potential temperature are calculated from existing data.  
171 Vertical resolution in the data varies over two of the long-term records. For example, vertical  
172 resolutions of 250 m are available for Boulder, CO launches from 1979 to 1989. Vertical  
173 resolutions of approximately 250 – 350 m (derived by recording one data point per minute) are  
174 available for Wallops Island, VA from 1970 to 1995. The remainder of the data have resolutions  
175 at or better than 100 m. Accounting for response time of the ozonesonde and the ascent rate of the  
176 balloon, the true vertical resolution of the O<sub>3</sub> measurements is approximately 100 – 150 m. For  
177 uniformity, data from all sondes were interpolated linearly to 100 m.

178       Ancillary meteorological data were added to assist the geophysical interpretation of the  
179 ozonesonde profile clusters. HYSPLIT (Hybrid Single Particle Lagrangian Integrated Trajectory;  
180 Draxler and Hess, 1997) back trajectories were computed starting at the time and location of each  
181 ozonesonde profile. The HYSPLIT trajectories were forced with NCEP/NCAR reanalysis (Kalnay  
182 et al., 1996), which is available globally from 1948 – present with 17 pressure levels and 2.5°x2.5°



horizontal resolution. Meteorological variables of temperature, potential vorticity (PV), and geopotential height were extracted at four levels (850, 700, 500, 250 hPa) from ERA-Interim (Dee et al., 2011) reanalysis. Mean sea level pressure (MSLP), total cloud cover, and 2 m temperature were also analyzed. ERA-Interim data are available globally from 1979 – present with 37 pressure levels and a horizontal resolution of  $\sim 0.7^\circ \times 0.7^\circ$ .

## 2.2. Self-Organizing Maps (SOM)

Presented here is a brief introduction to SOM clustering. Additional discussion of user-selectable SOM parameters appears in the Appendix. The Appendix also contains sensitivity tests and arguments that explain the selection of SOM over k-means for clustering, as well as the choice of SOM map size.

SOM, developed and described by Kohonen (1995), is an artificial neural network, that is, a network of “nodes” or “neurons” that learn from and are used to represent an input dataset. SOM is often used for data visualization or dimensional reduction of the original dataset (Liu et al., 2006 and references therein). In the application to CONUS O<sub>3</sub> mixing ratio profiles, SOM is employed primarily as a clustering algorithm.

The SOM algorithm is configured via a user-specified map (network) size and shape that dictates the number and relationship of the nodes that will represent the data. The map can be of any dimension, with 2-D maps (e.g. a 4x4 map of 16 nodes) preferred in recent related meteorological applications (e.g. Jensen et al., 2012; Nowotarski and Jensen, 2013). The initial values of the nodes, analogous to cluster centroids in k-means, can be obtained in a number of ways. Here, a Principal Component Analysis (PCA) decomposition of the input dataset yields a

subspace across which the initial nodes are distributed over a rectangular grid. This linear initialization approach is taken so as to cover as much of the input dataset variability as possible in the array of initial nodes. SOM nodes can also be initialized randomly as in k-means, although randomly initialized SOM maps converge more slowly and have larger error (Liu et al., 2006). Linear initialization also guarantees that the same map is consistently produced for a given set of inputs and is thus preferred here.

Once the SOM nodes are initialized, the SOM algorithm is executed on the input dataset in either the batch or sequential mode. These two iterative update modes result in similar SOMs, and the batch algorithm is much more computationally efficient (Vesanto et al., 2000), so batch is used in this study. The SOM algorithm clusters the input data (i.e. O<sub>3</sub> profiles) in such a way that each cluster is most similar to those holding adjacent positions in the map. This feature is utilized in the Results and Discussion section. This study uses code from the Matlab SOM Toolbox described in Vesanto et al. (2000), available for download from the Helsinki University of Technology, Finland (<http://www.cis.hut.fi/projects/somtoolbox/>). Further discussion of the SOM algorithm, comparisons and sensitivity tests with k-means clustering, and optimization of SOM geometry and topology for the CONUS sondes appear in Appendix A.

### 3. Results and Discussion

Average monthly O<sub>3</sub> mixing ratio profiles (surface – 12 km amsl) at the four CONUS locations are first presented (Figure 2). The least seasonal variability throughout the lower to mid-troposphere is observed at Trinidad Head, with O<sub>3</sub> mixing ratios generally averaging between 50 and 65 parts per billion by volume (ppbv) from 2 – 6 km throughout the year. Larger seasonal

variability in low to mid-troposphere  $O_3$  is observed at Huntsville and Wallops Island (50 – 75 ppbv). Higher summer  $O_3$  mixing ratios at the two latter sites reflect intermittent transport of photochemical pollution from upwind regional sources. Boulder exhibits less extreme seasonal variability. The small yearly range in low to mid-tropospheric  $O_3$  at Trinidad Head was first observed in the shorter (4.5 years) record examined by Newchurch et al. (2003). Trinidad Head also exhibits the lowest concentrations of near-surface and boundary layer  $O_3$  (<40 ppbv). Although Trinidad Head is often influenced by clean, marine air masses (Newchurch et al., 2003), the measurements may also be affected by local launch times that are 1 – 3 solar hours earlier than those of the other sites'. The majority of launches occur from 17 – 19 UTC at all locations, equating to 9 – 11 local standard time at the west coast Trinidad Head site.

The seasonal cycle in tropopause height, seen as sharp  $O_3$  increases near 10 – 12 km in Figure 2, displays a minimum in late winter/spring (MAM), and maximum in the fall (SON). In this study, tropopause height is calculated using the thermal lapse rate tropopause as defined by the World Meteorological Organization (WMO): the lowest altitude at which the temperature lapse rate increases to  $-2\text{ K km}^{-1}$  or less and persists for a depth of at least 2 km (WMO, 1957).

### 3.1. 3x3 SOM Cluster Results

In order to avoid the complexity of  $O_3$  gradients in the tropical tropopause layer (TTL), Jensen et al. (2012) ran SOM on profiles to 15 km amsl altitude. In contrast, we wish to capture variability in the tropopause altitude; therefore our SOM uses  $O_3$  mixing ratio data from the surface to 12 km amsl. At our CONUS locations, 12 km altitude is sufficient to include seasonal tropopause altitude variability in the SOM clusters, and encompasses most or all of the

252 troposphere. This altitude ceiling also prevents stratospheric O<sub>3</sub> mixing ratios from dominating  
253 the SOM clusters. The 3x3 SOM output (nine nodes/clusters) for Wallops Island is shown in  
254 Figure 3. Clusters of individual profiles (dark blue) corresponding to each SOM node (red) are  
255 plotted along with the entire dataset mean, 20<sup>th</sup>, and 80<sup>th</sup> percentile O<sub>3</sub> (cyan) for comparison.  
256 SOM nodes are labeled 1 – 9 and will be referred to by number when discussing characteristics of  
257 each O<sub>3</sub> profile cluster. Each SOM node is identical to the mean of its respective cluster. A major  
258 advantage of SOM over other clustering algorithms is adjacency of like nodes (e.g. 2 and 3, Figure  
259 3), and the separation of contrasting nodes (e.g. 3 and 7). This allows us to visualize subtle  
260 differences between the neighboring clusters of O<sub>3</sub> profiles, and distinguishes unique  
261 characteristics of nodes and groups of nodes through variation of specific features across the SOM  
262 map. For example, traversing nodes 1, 2, and 3 shows a lowering of the altitude of tropopause O<sub>3</sub>  
263 gradients; O<sub>3</sub> is well above the mean and 80<sup>th</sup> percentile O<sub>3</sub> near 8 – 12 km in these nodes.  
264 Likewise, a distinct rise in the amount of lower tropospheric O<sub>3</sub> from nodes 7 to 9 is observed.  
265 Node 7 contains profiles with O<sub>3</sub> below the 20<sup>th</sup> percentile and <50 ppbv through nearly the entire  
266 surface – 12 km profile. Node 8 contains near average O<sub>3</sub> in the low to mid-troposphere, and node  
267 9 exceeds 70 ppbv O<sub>3</sub> and remains above the 80<sup>th</sup> percentile through nearly the entire troposphere.

268         The percentage and total number of profiles corresponding to each node quantifies the  
269 frequency with which each O<sub>3</sub> profile type is observed, and which cluster(s) may be most  
270 representative of a site's typical O<sub>3</sub> profile. At Wallops Island, the top row of nodes 1 – 3 contains  
271 just 15% of all profiles, whereas the bottom row 7 – 9 contains 60%, and generally avoids  
272 tropopause O<sub>3</sub> gradients.

273         All nine SOM nodes from surface – 12 km O<sub>3</sub> mixing ratios for each of the four CONUS  
274 sites appear in Figure 4. The topological ordering and shapes of the SOM nodes are nearly

identical. This similar topological ordering from SOM clustering results from the linear initialization of SOM nodes. Differences in nodes 5 and 6 between Huntsville and the other sites is evident in Figure 4, presumably stemming from Huntsville's higher average tropopause and the subtropical-like characteristics noted in Newchurch et al. (2003) and Tilmes et al. (2012). Much like the Wallops Island SOM in Figure 3, there is a lowering of the altitude of O<sub>3</sub> gradients indicating the tropopause across nodes 1, 2, and 3. Nodes 1 – 3 represent 13 – 16% of each site's profiles. The increasing amount of lower tropospheric O<sub>3</sub> across nodes 7, 8, and 9, corresponding to 57 – 61% of profiles, is also prominent. Node 7 (Figure 4) represents a background state at every site, with O<sub>3</sub> averaging  $\leq 50$  ppbv throughout most of the troposphere, whereas node 9 contains polluted profiles with well above average low to mid-tropospheric O<sub>3</sub>.

### 3.2. Seasonal and Meteorological Influences on SOM Nodes

Given the dynamic and seasonal influences on O<sub>3</sub> profiles in the mid-latitudes, the SOM nodes at the CONUS ozonesonde locations are expected to yield information about, or correspond to seasonality, mid-latitude ridge and trough Rossby wave patterns, and O<sub>3</sub> regimes and column O<sub>3</sub> amounts. The number of profiles from each month in each SOM node is expressed as a percentage and is shown in Figure 5. Many of the clusters contain launches from only a few months. However, profiles from several nodes (e.g. 4 – 6) exhibit no distinct seasonality and are found throughout the year. The altitude of the tropopause O<sub>3</sub> gradient, and the photochemical O<sub>3</sub> season both contribute to the node/seasonality relationship. The lowest tropopause altitudes, predominantly found in nodes 1, 2, and 3, occur mainly in late winter and spring. During these seasons, increased latitudinal temperature gradients and synoptic-scale Rossby wave dynamics

298 cause large meanders in the polar jet, with associated lower tropopause altitude. At every site,  
299  $\geq 90\%$  of all profiles in node 3 occurred between January and May. Conversely, tropopause  $O_3$   
300 gradients are generally not found below 10 km in profiles corresponding to the bottom SOM row.  
301 Rather, the low to mid-tropospheric  $O_3$  increase across nodes 7 – 9 (~50 to 60 to 70 ppbv in Figure  
302 4) represents a sequence of increasing photochemical  $O_3$  pollution. The majority of profiles in  
303 node 7 occur in fall and winter, whereas  $>80\%$  of profiles in node 9 occur between May and August  
304 at all sites. Figures 4 and 5 show that there is reduced tropospheric  $O_3$  pollution year round at  
305 Trinidad Head compared to the other stations. Nearly 30% of node 7 profiles from Trinidad Head  
306 were launched in JJA. This is a far greater portion of summer launches in node 7 than for any  
307 other site. Fewer than 5% of launches in node 7 are from JJA at Huntsville and Wallops Island.  
308 Trinidad Head also includes the greatest percentage of its total launches in node 7 (31%) compared  
309 to other sites, and the fewest in the polluted node 9 (8%).

310 Many of the remaining SOM node clusters are difficult to explain through seasonality and  
311 tropopause height alone, and will require analysis of additional sources of data. Meteorological  
312 information examined from the radiosondes attached to the corresponding ozonesonde is used to  
313 infer influences on the SOM node  $O_3$  clusters. Radiosonde and ozonesonde data from SOM  
314 clusters are compared against monthly climatological means for each site. So that all CONUS  
315 locations can be considered together, an anomaly approach is taken.

316 Our approach to calculate anomalies from monthly climatology is as follows: 1) Calculate  
317 the climatology with 12 monthly means using a site's entire profile dataset for the variable of  
318 interest, 2) Calculate the variable of interest for every profile in each SOM node and compare to  
319 its corresponding monthly mean climatology (measurement – climatology), and 3) Average the  
320 results of all profiles within each SOM node. Average anomalies for each node at each site are

321 calculated. The result of applying this technique to the WMO lapse rate tropopause is shown in  
322 Figure 6a. Nodes 1 – 3 represent an incremental lowering of the tropopause altitude, with the  
323 lowest relative tropopause averaging over 4000 m lower than climatology in node 3 at Trinidad  
324 Head. Nodes 4 and 5 contain about average or slightly lower than average tropopause heights,  
325 whereas node 6 tropopause heights lie >1000 m below climatology at each site. Nodes 7 and 8  
326 appear to be related in that they contain similar tropopause height anomalies, with a slight increase  
327 in low to mid-tropospheric O<sub>3</sub> from node 7 to 8 as seasonality shifts from mainly fall-winter to  
328 spring-fall. The polluted, summertime node 9 tropopause altitudes are close to climatology, except  
329 at Huntsville, which averages 900 m higher. Huntsville's uniqueness is also displayed in node 5,  
330 a consequence of its more subtropical characteristics and distinct O<sub>3</sub> profiles and seasonality  
331 (Figures 4 and 5).

332 Anomalies of 500 hPa geopotential height and temperature (Figure 6b) are similar to the  
333 tropopause height anomalies (Figure 6a). Nodes 1 – 3 are associated with lower and progressively  
334 colder 500 hPa surfaces, suggesting that these profiles are influenced by Rossby wave troughs  
335 through most of the troposphere. Notably, Huntsville contains the largest 500 hPa height  
336 anomalies in node 3. Many of the 500 hPa surfaces in node 3 at Huntsville lie below 5.5 km.  
337 These are some of the lowest 500 hPa heights in the entire CONUS ozonesonde record. However,  
338 node 3 O<sub>3</sub> profiles represent only 2% of Huntsville's dataset. As in Figure 6a, nodes 4, 5, and 9  
339 lie close to climatological mean 500 hPa heights. The large-scale ridge pattern implied by the  
340 positive 500 hPa heights in node 7 is an indication of subtropical influence, and as a result, lower  
341 O<sub>3</sub> amounts. The 500 hPa heights and temperatures in node 6 are well below average and in line  
342 with its low tropopause height. With the evidence presented thus far, node 6 seems to be a  
343 miscellaneous cluster with highly variable O<sub>3</sub> profiles and unclear seasonality. However, node 6

contains  $\leq 5\%$  of data at each site. This cluster appears to be equivalent to the wastebin taxon in biology (e.g. Friedman et al., 2011).

### 3.3. Column O<sub>3</sub> Anomalies from Monthly Climatology

Using the same anomaly approach as in Figures 6a and 6b, integrated tropospheric column O<sub>3</sub> anomalies, calculated from the surface to the tropopause, for each SOM node (Figure 6c) are presented in Dobson Units (DU;  $2.69 \times 10^{16}$  molec. cm<sup>-2</sup>). The tropopause is defined using the same WMO lapse rate definition. Tropospheric column O<sub>3</sub> anomalies for nodes 1 – 3 display no pattern, unlike the meteorological anomalies in Figures 6a and 6b. There is no clear relationship between the low tropopause/500 hPa heights and tropospheric column amount. Average tropospheric column O<sub>3</sub> anomalies generally lie within a few DU or  $\pm 10\%$  of the climatology for most sites in nodes 1 – 3. However, distinct patterns in tropospheric O<sub>3</sub> amount are observed for nodes 7 – 9. The tropospheric O<sub>3</sub> increase from Figure 4 is prominent in tropospheric column anomalies in Figure 6c across node 7 (-4 to -8 DU; -10 to -20%), node 8 (+2 DU; +5%), and node 9 (+5 to +9 DU; +17 to +25%). Nodes 4 and 5 (Figure 6c) display near-climatological values, with tropospheric O<sub>3</sub> anomalies averaging  $\pm 2$  DU ( $\pm 5\%$ ). Stratospheric O<sub>3</sub> intrusions may contribute to the node 6 profiles, given the +4 – 10 DU (+10 – 25%) anomalies occurring in conjunction with low tropopause heights. Node 6 also contains the largest O<sub>3</sub> DU km<sup>-1</sup> values in the troposphere at all sites.

Total column O<sub>3</sub> is calculated from each sonde to investigate the synoptic meteorological influences on the integrated profile. Total column O<sub>3</sub> from the ozonesondes is derived by first integrating the O<sub>3</sub> profile from surface to balloon burst or 10 hPa, whichever is higher in pressure



(lower in altitude). The 10 hPa cut off has been shown to reduce errors in the total column O<sub>3</sub> calculation resulting from increasing measurement uncertainties in the mid-stratosphere (e.g. Stauffer et al., 2014). The McPeters and Labow (2012) above-balloon burst O<sub>3</sub> column climatology is then added to the ozonesonde column O<sub>3</sub> amount, yielding a total column O<sub>3</sub> amount. The McPeters and Labow (2012) O<sub>3</sub> climatology is based on a combination of ozonesonde and Aura Microwave Limb Sounder (MLS) climatology. Profiles that did not reach 30 hPa were discarded.

The resulting total column O<sub>3</sub> anomalies in Figure 6d reflect the tropopause height anomalies in Figure 6a. Nodes 1 – 3 contain increasing total column O<sub>3</sub> corresponding to the lowering tropopause, yielding a deeper stratosphere. Node 3 profile columns frequently exceed 400 DU, representing a ~55 – 75 DU (~15 – 25%) increase in total column O<sub>3</sub> over climatology. An increase of about 20 – 30 DU (10%) above average O<sub>3</sub> appears in the highly variable O<sub>3</sub> profiles in node 6. Nodes 4, 5, 8, and 9 contain near-average total column O<sub>3</sub> within ±5% of climatology. Node 7 is the only cluster with notably low total column O<sub>3</sub>, 20 – 35 DU below the climatological average.

### **3.4. Meteorological Interpretations**

The meteorological and seasonal influences on nodes 1 – 3 (winter/spring, low tropopause O<sub>3</sub> profiles) and 7 (fall/winter, high tropopause/500 hPa heights, subtropical influence), are obvious. However, ancillary data are required to interpret the remaining SOM nodes. A contoured heat map of HYSPLIT back trajectories ending at 4 km provides a summary of source regions for all SOM nodes and sites (Figure 7). The 4 km altitude was chosen because this altitude is typically

located in the free troposphere, well above effects from boundary layer processes, yet low enough to avoid the largely zonal winds at higher altitude that result from thermal wind balance. Distinct cyclonic curvature in the trajectories is evident in nodes 2, 3, and 6 in Figure 7, confirming that large-scale troughs (e.g. 500 hPa heights, Figure 6b) are the driving force behind the profiles in those clusters. In node 7 at Huntsville and Wallops Island, anti-cyclonic curvature and a source region to the southwest illustrate the previously noted subtropical influences; node 4 contains similar trajectories. Trajectories from other nodes are mostly zonal.

Meteorological variables from ERA-Interim reanalysis were analyzed to further explore the origins of nodes 4 – 6, 8, and 9. Each node is evaluated individually using meteorological anomalies calculated with the same methodology used for Figure 6. We note that the beginning of the Wallops Island record is not covered by the ERA-Interim dataset (1979 – present), but this fact is not expected to influence the following results as a large sample size remains. While some of the variables examined could simply be extracted from the sonde data, it is more useful to examine 3-D meteorological fields to aid geophysical interpretation of the remaining nodes.

Node 4 exhibits negative PV anomalies at 250 hPa (Figure 8), indicating reduced stratospheric influence at upper levels. The anomalies, on the order of -0.3 to -1.0 potential vorticity units (PVU), are observed at all sites. Positive anomalies of MSLP and geopotential height at 850 and 700 hPa at all sites (see Supplementary Figures S1 and S2), and the southwesterly trajectory tendency at Huntsville and Wallops Island (Figure 7), support the hypothesis that node 4 profiles are indeed influenced by subtropical air, leading to slightly below-average tropospheric O<sub>3</sub> amounts. Quantitative values of the meteorological anomalies (Figure 6a-b), however, make it apparent that the degree to which subtropical air affects node 4 profiles is much less than for node 7.

413       Node 5 profiles exhibit slightly above average total and tropospheric column O<sub>3</sub>, and  
414 slightly below average tropopause and 500 hPa geopotential heights (Figure 6). ERA-Interim 250  
415 hPa geopotential height anomalies (Figure 9) exhibit an upper-level trough influence on these  
416 profiles with negative anomalies (except Huntsville) of -35 to -90 m. The disparity between  
417 Huntsville and the other sites is evident in the 250 hPa height anomalies, consistent with the node  
418 5 differences in SOM profile shape in Figure 4 at 10 – 12 km. The lower 250 hPa heights reflect  
419 a correspondingly lower tropopause height which increases O<sub>3</sub> at this level, an effect not as evident  
420 in node 5 Huntsville profiles as at other sites. The remainder of the ERA-Interim pressure level  
421 and surface data are void of anomalies in node 5; only altitudes near the tropopause show  
422 appreciable meteorological influence on the O<sub>3</sub> profiles.

423       Node 6 profiles are hypothesized to be influenced by STE, given the low 500 hPa and  
424 tropopause heights (Figure 6a-b), well-above average tropospheric column O<sub>3</sub> (Figure 6c), and  
425 cyclonic-curving back trajectories (Figure 7). Positive maxima of 500 hPa PV anomalies (Figure  
426 10; +0.1 to +0.4 PVU) centered near each site indicate stratospheric influence in the mid-levels of  
427 node 6 profiles. Though as node 6 members represent an assortment of profiles, many contain  
428 layers of high O<sub>3</sub> mixing ratios from stratospheric origins in the mid-troposphere.

429       Node 8 profiles contain positive tropopause and 500 hPa height anomalies similar to node  
430 7 profiles (Figure 6), but contain near-average tropospheric column O<sub>3</sub> amounts. The 500 hPa  
431 geopotential height anomalies derived from ERA-Interim are shown in Figure 11. The 500 hPa  
432 geopotential height anomalies derived from ERA-Interim (see Figure 11) differ only slightly from  
433 those derived from sonde data. A clear trough-ridge structure is visible in the average 500 hPa  
434 geopotential height anomalies of node 8 throughout all CONUS sites. Node 8 profiles have  
435 positive geopotential height and temperature anomalies through all four (850, 700, 500, 250 hPa)

extracted pressure levels. This is also true for node 7 profiles. In terms of meteorological anomalies, node 8 is nearly identical to node 7. Thus, the dichotomy in seasonality (Figure 5) of node 7 and 8 is likely the major driver behind the O<sub>3</sub> profile differences in these two clusters.

Node 9 profiles have fewer defining meteorological characteristics. The most distinct features are observed in the ERA-Interim temperature anomalies, especially at 2 m (Figure 12). All sites but Trinidad Head are anomalously warm (+0.7 to +1.5 °C) near the surface in the largely summertime node 9 profiles. Trinidad Head, which is rarely polluted near the surface and contains relatively few profiles in node 9, averages 1 °C cooler than climatology at 2 m. This temperature behavior holds true through most of the troposphere, with Boulder, Huntsville, and Wallops Island being warmer than normal at 2 m, 850, 700, and 500 hPa, and Trinidad Head being cooler than normal at these levels. Additionally, Trinidad Head exhibits positive PV anomalies at 500 hPa in node 9 profiles. Processes leading to enhanced tropospheric O<sub>3</sub> amounts at Trinidad Head, such as STE and transport of pollution across the Pacific Ocean, are different from the other three sites. Other than anomalous temperatures at all sites and PV in the mid-levels at Trinidad Head, there is a lack of significant dynamic meteorological forcing in node 9. Therefore, node 9 is hypothesized to result from transported pollution during the high sun angle summer months, facilitating photochemical production in the troposphere and the resulting O<sub>3</sub> profiles.

### **3.5. O<sub>3</sub> Profile Anomalies from Monthly Climatology**

SOM clusters show characteristics based on meteorological measurements and reanalysis, seasonality, and O<sub>3</sub> column amounts. The next step is to evaluate how closely the monthly O<sub>3</sub> climatology describes the vertical O<sub>3</sub> profiles in the SOM node clusters. The average difference

between profiles from each node and their respective monthly O<sub>3</sub> mixing ratio climatology is calculated. Results for all nodes at each site are presented in terms of O<sub>3</sub> mixing ratio anomalies in ppbv (Figure 13). Not surprisingly, given their low tropopause heights, nodes 1, 2, and 3 exceed monthly climatological O<sub>3</sub> by over 100 ppbv, and in many cases more than double the climatological O<sub>3</sub> from 8 to 12 km. Extreme O<sub>3</sub> increases above climatology of over +75 ppbv also appear at all sites in node 6. Ozone anomalies in node 6 retreat above 10 km at Boulder and Huntsville, accounting for the complex O<sub>3</sub> profile shapes and stratospheric intrusion layers in that cluster. Nodes 4 and 5 differ by only a few ppbv from climatology, typically within  $\pm 3$  ppbv in the low to mid-troposphere. The largest deviations from climatological O<sub>3</sub> in nodes 4 and 5 occur above 10 km, coincident with the PV and geopotential height anomalies evident in Figures 8 and 9. Even discounting variations in the tropopause region, climatological O<sub>3</sub> averages may fail to describe a large percentage of the O<sub>3</sub> profiles at the CONUS sites. At 6 km at all sites, node 7 profiles on average fall more than 6 ppbv below monthly climatology, in conjunction with tropospheric column O<sub>3</sub> amounts that are 10 – 20% below normal (Figure 6c). Conversely, node 9 profiles lie well above climatology, exceeding it by over +7 ppbv at all sites, and up to +10 ppbv at Wallops Island and Trinidad Head. Nodes 7 and 9 correspond to 36 – 39% of all profiles at each of the ozonesonde sites. Clearly, monthly O<sub>3</sub> climatologies do not adequately describe the variability of CONUS O<sub>3</sub> observations, either in terms of profile shape, or column O<sub>3</sub> amount. Use of nine SOM clusters has more accurately captured the distribution of these O<sub>3</sub> profile data sets than monthly averages, particularly near the highly variable tropopause altitude.

Based on findings from Figures 6a, 6c, and 6d, one expects a relationship between the tropopause height and tropospheric column O<sub>3</sub>. Figure 14 shows how that relationship may change given the O<sub>3</sub> anomalies observed in Figures 6c and 13. Figure 14 presents a scatterplot and least-

squares fits of tropopause height and tropospheric column  $O_3$ . Except for node 7 profiles, each SOM node results in a similar correlation, or contains few enough profiles so as to not affect the overall dataset correlation. In fact, excluding node 7 profiles greatly increases the total dataset correlation coefficient between tropopause height and tropospheric column  $O_3$  at all sites (Boulder,  $r = 0.59$  to  $0.73$ ; Huntsville,  $r = 0.57$  to  $0.69$ ; Trinidad Head,  $r = 0.44$  to  $0.65$ ; Wallops Island,  $r = 0.53$  to  $0.71$ ). The low tropospheric  $O_3$  amount (lowest average of all nodes) and high tropopause heights found in node 7 represent a separate regime (red line in Figure 14) signaled by a displacement in the relationship between the two variables compared to the rest of the dataset. SOM node 7 profiles contain the lowest  $O_3$  DU  $km^{-1}$  value of all nodes in the troposphere. The typical tropopause height/tropospheric column  $O_3$  relationship observed in 70 – 80% of CONUS profiles is adjusted when fall and winter profiles contain positive tropopause and 500 hPa height anomalies.

A summarizing table (Table 2) is provided to outline the meteorological and  $O_3$  characteristics of the SOM nodes at each site.

#### 4. Conclusions

The application of SOM to 4530 CONUS  $O_3$  profiles led to clustering that is primarily based on two main factors: 1) The altitude of the tropopause  $O_3$  gradient, and 2) The amount of  $O_3$  in the low to mid-troposphere. Though profiles in some SOM clusters were mostly confined to a few months in the year, several exhibited no distinct seasonality, indicating more than just temporal effects on  $O_3$  profile variability over CONUS. The top row of nodes 1 – 3 (13 – 16% of CONUS profiles) at each site represented an incremental lowering of the tropopause  $O_3$  gradient.

The nodes were associated with synoptic-scale troughs, and contained double the climatological O<sub>3</sub> amount from 8 – 12 km. Thus, capturing the variability in tropopause height is vital for reproducing the significant day-to-day changes in O<sub>3</sub> profile shape at these CONUS sites.

Challenges in characterizing CONUS O<sub>3</sub> variability go beyond knowledge of the tropopause height. Nodes 7 and 9 displayed the largest deviations from climatological O<sub>3</sub> in the low to mid-troposphere. Ozone in nodes 7 and 9 was generally beyond  $\pm 6$  ppbv from 6 km climatological O<sub>3</sub>, but up to +10 ppbv (+25% tropospheric column O<sub>3</sub>) in node 9 at Trinidad Head and Wallops Island. Inclusion of node 7 profiles, which represent a different regime in the tropopause height/tropospheric column O<sub>3</sub> relationship, greatly reduced correlation coefficients between tropopause height and tropospheric column O<sub>3</sub> for the entire dataset. Nodes 7 and 9 contained nearly 40% of CONUS O<sub>3</sub> profiles. Understanding the large-scale conditions outlined here that lead to clean tropospheric O<sub>3</sub> profiles of subtropical origins (node 7), and summertime tropospheric pollution events (node 9) is key to better design of chemical models and satellite algorithms.

Although SOM nodes are explained by large-scale meteorological conditions, we will explore SOM connections for locations and periods with more chemical measurements.

Commented [AMT1]: took out one 'large'

The finding that simple time means are inadequate for describing the complexity of individual O<sub>3</sub> profiles is particularly true near the tropopause at the CONUS sites addressed here. The diversity of CONUS O<sub>3</sub> profile shapes is more appropriately expressed using nine SOM clusters than by using 12 monthly averages. In the mid-latitudes, O<sub>3</sub> profile evolution is highly dependent upon trough and ridge systems associated with large-scale Rossby waves, pollution transport, and the influence of subtropical air, all of which cause appreciable changes over short time scales. SOM graphically depicts the contributions to this variability in the tropospheric O<sub>3</sub>

profile. SOM provides exceptional insights into the seasonality, or lack thereof, of observed profile shapes and the frequency of extreme, dynamic-induced changes at the tropopause level observed in SOM nodes 1 – 3 in this paper. SOM also distinguishes O<sub>3</sub> profiles with very low and high (nodes 7 and 9) tropospheric O<sub>3</sub> amounts, useful for quantifying typical baseline and polluted O<sub>3</sub> levels.

## Appendix A

This Appendix provides explanations of the k-means clustering algorithm and details of the SOM clustering user-selectable settings. Sensitivity tests on the user-selectable SOM settings are then compared to the output with k-means, using both randomly-generated and real ozonesonde O<sub>3</sub> mixing ratio profiles. Results from these tests are used to justify our choice of clustering algorithm ultimately used in this study, the 3x3 SOM map with nine nodes/clusters.

### k-means

The k-means algorithm partitions an input dataset into a user-defined number ( $k$ ) of clusters. Cluster centroids are initialized through random selection of  $k$  vectors from the input dataset. Each remaining input vector is then assigned to the closest (in Euclidean distance) centroid. Finally, the centroid of each cluster is updated:

$$\mathbf{m}_i(t+1) = \frac{1}{n_i} \sum_{j \in n_i} \mathbf{x}_j \quad (\text{A1})$$



551 where  $\mathbf{m}$  is the  $i$ th cluster centroid,  $\mathbf{x}$  is the  $j$ th data vector belonging to the  $i$ th cluster,  $n$  is the  
552 number of data vectors belonging to the  $i$ th cluster, and  $t$  is the iteration. The new centroid is thus  
553 the average of the previous centroid's assigned input vectors. All of the data vectors are then  
554 reassigned to clusters based on these new centroids. This process is repeated until there are no  
555 new vector assignments — the algorithm has converged, with the input data separated into  $k$   
556 exclusive clusters. Defining the centroids as the mean of the corresponding data vectors guarantees  
557 that the average Euclidean distance between a cluster's centroid and its member vectors is  
558 minimized.

559

#### 560 **The Batch SOM Algorithm**

561

562 In the batch SOM algorithm, each vector in the dataset is grouped with its closest (in  
563 Euclidean distance) initial SOM node, called the best matching unit (BMU). The batch SOM  
564 equation is then applied to the dataset and can be repeated thousands of times (each repeat is called  
565 an epoch) if necessary to converge to a final map. The batch equation is as follows:

566

$$567 \quad \mathbf{m}_i(t+1) = \sum_{j=1}^M n_j h_{ij}(t) \bar{\mathbf{x}}_j / \sum_{j=1}^M n_j h_{ij}(t) \quad (\text{A2})$$

568

569 where  $\mathbf{m}$  is the  $i$ th of  $M$  total nodes,  $n$  is the number of vectors for which node  $j$  is the BMU,  $\bar{\mathbf{x}}$  is  
570 the mean of the vectors for which node  $j$  is the BMU,  $h$  is the value of the neighborhood function  
571 (dependent on factors discussed below), and  $t$  is the epoch. This equation is repeatedly calculated  
572 for the user-defined number of epochs. Essentially, each  $\mathbf{m}_i$  node is updated with the mean of its

573 member vectors  $\bar{\mathbf{x}}_j$  when  $i = j$ , and  $\bar{\mathbf{x}}_j$  multiplied by the neighborhood function, a value from 0 to  
574 1, when  $i \neq j$ .

575

## 576 **The SOM Neighborhood Function**

577

578 The neighborhood function distinguishes the batch trained SOM from the k-means  
579 algorithm. This function allows updating node  $\mathbf{m}_i$  to learn from nearby nodes' member vectors in  
580 addition to its own. Typically in SOM, the neighborhood function value decreases gradually with  
581 increasing distance between nodes  $i$  and  $j$ . This causes node  $\mathbf{m}_i$  to learn less from neighboring  
582 nodes' member vectors than its own, but more than is the case with k-means where such inter-  
583 cluster learning does not occur at all. The addition of inter-cluster learning leads to a topographical  
584 ordering of SOM nodes that is absent in k-means clusters. However, in equation (A2), if the  
585 neighborhood function is 1 only when  $i = j$ , and 0 otherwise, the SOM learning becomes identical  
586 to k-means. In that limiting case, each node  $\mathbf{m}_i$  is updated only with its own member vectors, there  
587 is no neighborhood node learning, and the nodes are independent of each other.

588 The neighborhood function depends on the Euclidean distance between the updating node  
589  $\mathbf{m}_i$ , the current node iteration  $j$ , and the user-defined neighborhood radius. The distance between  
590 nodes is 0 when  $i = j$ , ranging to  $\sqrt{32}$  in a 4x4 map, for example. The neighborhood radius is  
591 reduced linearly with epoch so the neighborhood function value decays, diminishing the  
592 neighborhood learning and allowing the map solution to converge. We explore the effects of four  
593 neighborhood functions, each available as an option in the Matlab SOM Toolbox (Vesanto et al.,  
594 2000; Liu et al., 2006), as follows:

595

$$\begin{aligned}
596 \quad h_{ij}(t) = & \begin{cases} \exp\left(-\frac{d_{ij}^2}{2r_t^2}\right) & \text{Gaussian} \\ \exp\left(-\frac{d_{ij}^2}{2r_t^2}\right)F(r_t^2 - d_{ij}^2) & \text{Cut Gauss} \\ F(r_t^2 - d_{ij}^2) & \text{Bubble} \\ \left[\frac{1-d_{ij}^2}{r_t^2}\right]F(r_t^2 - d_{ij}^2) & \text{Epanechnikov (Ep)} \end{cases} \quad (A3)
\end{aligned}$$

597  
598 Here,  $r$  is the neighborhood radius at epoch  $t$ ,  $d$  is the distance between nodes  $i$  and  $j$ , and  $F(x)$  is a  
599 step function with a value of 1 if  $x \geq 0$ , and 0 otherwise. The basic geometry of these functions is  
600 found in Vesanto et al. (2000). The Gaussian function decays to a non-zero value as distance  
601 between nodes increases. The Cut Gauss, Bubble, and Ep functions are zero once the node distance  
602  $d$  is greater than the neighborhood radius  $r$ .

#### 603 604 **SOM Neighborhood Functions Test**

605  
606 The user's choice of map size/cluster number and SOM neighborhood function can have  
607 significant impacts on the amount of data assigned to each cluster and the quality of fit between a  
608 node/centroid and its member data. A sensitivity analysis is conducted to determine user-chosen  
609 parameter settings that produce the most effective O<sub>3</sub> profile clusters from SOM or k-means.  
610 Neighborhood functions will be evaluated first, followed by map size/cluster number.

611 To compare the performance of k-means and the four SOM neighborhood functions, each  
612 algorithm was applied to an identical, random, 1000 point, 2-D dataset using a 3x3 SOM map and,  
613 equivalently,  $k = 9$  clusters (Figure A1). For this test, the SOM neighborhood radius decreased  
614 linearly from 3 to 1 over 100 epochs. The k-means algorithm was repeated to convergence and  
615 was initialized identically to the SOMs (via PCA) to avoid the stochastic outcomes that typically

616 result from random initialization (a random k-means initialization will be considered in other  
617 tests). Figure A1 represents how neighborhood functions organize and cluster a randomly  
618 generated set of 2-D data. Given the use of neighborhood learning, the nodes in SOM depend  
619 upon others' member vectors, in contrast with the independent and arbitrarily organized clusters  
620 in k-means. Consistent ordering of SOM nodes, regardless of neighborhood function, is displayed  
621 by the SOM node number labels in Figure A1.

622         Clustering often seeks to maximize the distance between centroids/nodes to better  
623 distinguish signals in the dataset. The Ep function SOM nodes (Figure A1; top left, large color  
624 dots) converge most similarly to the k-means algorithm (black diamonds) because of the Ep  
625 neighborhood functions' sharp decrease with increasing distance between nodes; each node is less  
626 dependent on non-member vectors than other functions. This yields the greatest distances among  
627 nodes among the neighborhood functions. Gaussian, the slowest decaying function with node  
628 distance, yields nodes that cluster close together near the overall dataset mean, greatly contrasting  
629 the independent and farthest separated k-means centroids. The varied clustering resulting from  
630 use of different neighborhood functions also causes disparity in performance as measured by error  
631 metrics, as explained in the next section.

632

633 **Error Metrics**

634

635         There are two standard measures of SOM error: Quantization Error (QE) and  
636 Topographical Error (TE). Figure A2 shows QE, the average Euclidean distance between a  
637 node/centroid and its respective member vectors, for both 3x3 SOM and k-means (both random  
638 and PCA-initialized are included). SOM and k-means are performed on the combined CONUS

639 sites' O<sub>3</sub> mixing ratio profile datasets. The altitude range covers surface – 12 km amsl as  
640 throughout the paper.

641 The four-site average QE value (small values are desired) is shown as a function of epoch  
642 in Figure A2. To remain consistent with SOM settings used in the body of the paper, QE values  
643 up to 1000 epochs are presented (note that k-means converges to its solution long before 1000  
644 iterations). Given the distribution of nodes for each neighborhood function in Figure A1, the  
645 results are not surprising. The Ep function mimics both k-means runs, which by definition  
646 minimize the QE metric for clusters. Because the Gaussian function clusters tend toward the  
647 overall mean, the fits between its nodes and member data are the worst of the six options shown  
648 in Figure A2. The Bubble and Cut Gauss functions fall between these two extremes.

649 TE provides a measure of how well the SOM map fits the data manifold. TE is the fraction  
650 of input data vectors whose BMU is *not* adjacent to its second closest node (Figure A3; same SOM  
651 settings as Figure A2), with smaller percentages generally indicating better organization. This  
652 metric quantifies a major advantage of SOM over k-means. The organization of clusters by SOM  
653 provides superior data visualization (Figures 3 and 4). Because k-means clusters are randomly  
654 ordered and unorganized, TE is not relevant to them. In Figure A3, the Gaussian neighborhood  
655 function yields the lowest TE, indicating adjacent nodes are most alike with that function. TE is  
656 highest for the Ep function, which may be a result of its more independent, and thus more unique,  
657 nodes. Still, the TE metric varies by only a few percent between neighborhood functions, and the  
658 errors show well-ordered maps in any case. Given the small variation in TE between neighborhood  
659 functions, and the improved QE and uniqueness performance of Ep compared to other  
660 neighborhood functions, we choose the Ep function to further compare SOM against k-means.

661

## Cluster Number/SOM Map Size

The final sensitivity test evaluates randomly- and PCA-initialized k-means, and SOM using the Ep neighborhood function. A balance is sought between the number of profiles assigned to each cluster, and the total number of clusters. The choice of number of clusters must be enough to capture the variability in the dataset, and each cluster must contain enough cases to sufficiently describe geophysical meaning for each cluster. The percentage of profiles in both the most and least populous clusters for each case is chosen to provide a measure of this balance. For each ozonesonde location, varying numbers of clusters and SOM nodes are analyzed to evaluate cluster membership (Table A1). The SOMs were run to 1000 epochs with the same settings as prior SOM tests. Small  $k$  and SOM map sizes result in highly populated clusters. Several of the SOM and k-means clusters contain over half the data in the 2x2 SOM and  $k = 4$  k-means solutions. At all CONUS sites, the two most populated clusters in the 2x2 SOM/ $k = 4$  solutions contain ~80% of all profiles, making characterization and interpretation of those clusters difficult. This statement is supported by closer inspection of differences between profile membership in the 2x2 and 3x3 SOM options (Table S1). The CONUS O<sub>3</sub> profile 2x2 SOM nodes are combinations of specific 3x3 SOM nodes, masking the unique meteorological conditions characteristic of many (particularly nodes 4, 5, 6, and 8) 3x3 SOM nodes found in the body of this paper.

As the map size and cluster number increases, membership of the least populous SOM and k-means clusters drops precipitously. The k-means centroids appear to be affected by outlier profiles, yielding several one-member clusters when  $k = 16$ . Presumably, this results from a lack of neighborhood learning. Even when  $k = 9$ , the least populous k-means cluster contains 1% or less of O<sub>3</sub> profiles in several cases. Considering the excessively large cluster membership in the

685 2x2 SOM/ $k = 4$  k-means, and the lack of profiles associated with nodes and centroids in the 4x4  
686 SOM/ $k = 16$  or 9 k-means, the 3x3 SOM with the Ep neighborhood function appears to be optimum  
687 for examining O<sub>3</sub> profile clustering at each site.  
688

## Acknowledgements

Funding for this project was provided by the following NASA grants: NNG05G062G, NNX10AR39G, NNX11AQ44G, and NNX12AF05G. The continued operation of CONUS ozonesonde stations are the combined efforts of many institutions and individuals: Boulder, CO and Trinidad Head, CA: Samuel Oltmans and Bryan Johnson (NOAA ESRL GMD); Huntsville, AL: Michael Newchurch (University of Alabama – Huntsville); Wallops Island, VA: Frank Schmidlin and E. Thomas Northam (NASA/Wallops Flight Facility). Thanks to the World Ozone and Ultraviolet Radiation Data Centre (WOUDC) for continued availability of ozonesonde datasets. Thanks also to Bryan Johnson for providing high resolution profile data from 1979 – 1989 for the Boulder, CO station. Thanks to Anders Jensen (Penn State University) for initial assistance with SOM. WOUDC data accessed at: <ftp://ftp.tor.ec.gc.ca/pub/woudc/>. NOAA ESRL GMD data accessed at: <ftp://ftp.cmdl.noaa.gov/data/ozwv/Ozonesonde/>. ERA-Interim reanalysis data accessed at: <http://rda.ucar.edu/datasets/ds627.0/>. NCEP/NCAR reanalysis data accessed at: <ftp://ftp.cdc.noaa.gov/>. This paper is the basis for a chapter in the first author's PhD thesis. The authors also thank the Editor and three anonymous reviewers for suggestions that improved this manuscript.



## References

- Considine, D. B., J. A. Logan, and M. A. Olsen (2008), Evaluation of near-tropopause ozone distributions in the Global Modeling Initiative combined stratosphere/troposphere model with ozonesonde data, *Atmos. Chem. Phys.*, 8, 2365-2385, doi:10.5194/acp-8-2365-2008.
- Cooper, O. R., et al. (2011), Measurement of western U.S. baseline ozone from the surface to the tropopause and assessment of downwind impact regions, *J. Geophys. Res.*, 116, D00V03, doi:10.1029/2011JD016095.
- Danielsen, E. F. (1968), Stratospheric-tropospheric exchange based on radioactivity, ozone and potential vorticity, *J. Atmos. Sci.*, 25, 502-518, doi:http://dx.doi.org/10.1175/1520-0469(1968)025<0502:STEBOR>2.0.CO;2.
- Dee, D. P., et al. (2011), The ERA-Interim reanalysis: configuration and performance of the data assimilation system, *Q. J. R. Meteorol. Soc.*, 137, 553-597, doi:10.1002/qj.828.
- Diab, R. D., A. M. Thompson, K. Mari, L. Ramsay, and G. J. R. Coetzee (2004), Tropospheric ozone climatology over Irene, South Africa, from 1990 to 1994 and 1998 to 2002, *J. Geophys. Res.*, 109, D20301, doi:10.1029/2004JD004793.
- Draxler, R. R., and G. D. Hess (1997), Description of the HYSPLIT\_4 modeling system, NOAA Tech. Memo, ERL ARL-224, NOAA Air Resources Laboratory, Silver Spring, MD, 24 pp.
- Friedman, M., and M. D. Brazeau (2011), Sequences, stratigraphy and scenarios: What can we say about the fossil record of the earliest tetrapods?, *P. Roy. Soc. Edinb. B.*, 278, 432-439, doi:10.1098/rspb.2010.1321.
- Hewitson, B. C., and R. G. Crane (2002), Self-organizing maps: Applications to synoptic climatology, *Clim. Res.*, 22, 13-26, doi:10.3354/cr022013.

730 Holton, J. R., P. H. Haynes, M. E. McIntyre, A. R. Douglass, R. B. Rood, and L. Pfister (1995),  
 731 Stratosphere-troposphere exchange, *Rev. Geophys.*, 33(4), 403-439,  
 732 doi:10.1029/95RG02097.  
 733 Hong, Y., K. Hsu, S. Sorooshian, and X. Gao (2004), Precipitation estimation from remotely  
 734 sensed imagery using an artificial neural network cloud classification system, *J. Appl.*  
 735 *Meteorol.*, 43, 1834-1853, doi:10.1175/JAM2173.  
 736 Jensen, A. A., A. M. Thompson, and F. J. Schmidlin (2012), Classification of Ascension Island  
 737 and Natal ozonesondes using self-organizing maps, *J. Geophys. Res.*, 117, D04302, doi:  
 738 10.1029/2011JD016573.  
 739 Kalnay, E., et al. (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Amer. Meteorol.*  
 740 *Soc.*, 77, 437-471, doi:http://dx.doi.org/10.1175/1520-  
 741 0477(1996)077<0437:TNYRP>2.0.CO;2.  
 742 Kohonen, T. (1995), The Basic SOM, in *Self-Organizing Maps*, pp. 77-130, Springer, New  
 743 York.  
 744 Komhyr, W. D. (1969), Electrochemical concentration cells for gas analysis, *Ann. Geophys.*, 25,  
 745 203-210.  
 746 Komhyr, W. D., R. A. Barnes, G. B. Brothers, J. A. Lathrop, and D. P. Opperman (1995),  
 747 Electrochemical concentration cell ozonesonde performance evaluation during STOIC  
 748 1989, *J. Geophys. Res.*, 100(D5), 9231-9244, doi:10.1029/94JD02175.  
 749 Lamarque, J.-F., et al. (2012), CAM-chem: description and evaluation of interactive atmospheric  
 750 chemistry in the Community Earth System Model, *Geosci. Model Dev.*, 5, 369-411,  
 751 doi:10.5194/gmd-5-369-2012.

752 Lin, M., A. M. Fiore, O. R. Cooper, L. W. Horowitz, A. O. Langford, H. Levy II, B. J. Johnson,  
 753 V. Naik, S. J. Oltmans, and C. J. Senff (2012), Springtime high surface ozone events over  
 754 the western United States: Quantifying the role of stratospheric intrusions, *J. Geophys.*  
 755 *Res.*, 117, D00V22, doi:10.1029/2012JD018151.

756 Liu, Y., R. H. Weisberg, and C. N. K. Mooers (2006), Performance evaluation of the self-  
 757 organizing map for feature extraction, *J. Geophys. Res.*, 111, C05018,  
 758 doi:10.1029/2005JC003117.

759 Logan, J. A. (1985), Tropospheric ozone: Seasonal behavior, trends, and anthropogenic  
 760 influence, *J. Geophys. Res.*, 90(D6), 10463–10482, doi:10.1029/JD090iD06p10463.

761 Logan, J. A. (1994), Trends in the vertical distribution of ozone: An analysis of ozonesonde data,  
 762 *J. Geophys. Res.*, 99(D12), 25553–25585, doi:10.1029/94JD02333.

763 Logan, J. A., et al. (1999), Trends in the vertical distribution of ozone: A comparison of two  
 764 analyses of ozonesonde data, *J. Geophys. Res.*, 104(D21), 26373–26399,  
 765 doi:10.1029/1999JD900300.

766 Logan, J. A., et al. (2012), Changes in ozone over Europe: Analysis of ozone measurements from  
 767 sondes, regular aircraft (MOZAIC) and alpine surface sites, *J. Geophys. Res.*, 117,  
 768 D09301, doi:10.1029/2011JD016952.

769 McPeters, R. D., G. J. Labow, and B. J. Johnson (1997), A satellite-derived ozone climatology  
 770 for balloonsonde estimation of total column ozone, *J. Geophys. Res.*, 102, D7, 8875–  
 771 8885, doi:10.1029/96JD02977.

772 McPeters, R. D., and G. J. Labow (2012), Climatology 2011: An MLS and sonde derived ozone  
 773 climatology for satellite retrieval algorithms, *J. Geophys. Res.*, 117, D10303,  
 774 doi:10.1029/2011JD017006.

775 Newchurch, M. J., M. A. Ayoub, S. Oltmans, B. Johnson, and F. J. Schmidlin (2003), Vertical  
 776 distribution of ozone at four sites in the United States, *J. Geophys. Res.*, 108(D1), 4031,  
 777 doi:10.1029/2002JD002059.  
 778 Nowotarski, C. J., and A. A. Jensen (2013), Classifying proximity soundings with self-  
 779 organizing maps toward improving supercell and tornado forecasting, *Wea. Forecasting*,  
 780 28, 783–801, doi:10.1175/WAF-D-12-00125.1.  
 781 Oltmans, S. J., et al. (2006), Long-term changes in tropospheric ozone, *Atmos. Environ.*, 40(17),  
 782 3156–3173, doi:10.1016/j.atmosenv.2006.01.029.  
 783 Oltmans, S. J., et al. (2013), Recent tropospheric ozone changes – A pattern dominated by slow  
 784 or no growth, *Atmos. Environ.*, 67, 331–351, doi:10.1016/j.atmosenv.2012.10.057.  
 785 Rao, T. N., S. Kirkwood, J. Arvelius, P. von der Gathen, and R. Kivi (2003), Climatology of  
 786 UTLS ozone and the ratio of ozone and potential vorticity over northern Europe, *J.*  
 787 *Geophys. Res.*, 108(D22), 4703, doi:10.1029/2003JD003860.  
 788 Stauffer, R. M., G. A. Morris, A. M. Thompson, E. Joseph, G. J. R. Coetzee, and N. R. Nalli  
 789 (2014), Propagation of radiosonde pressure sensor errors to ozonesonde measurements,  
 790 *Atmos. Meas. Tech.*, 7, 65–79, doi:10.5194/amt-7-65-2014.  
 791 Stevenson, D. S., et al. (2006), Multimodel ensemble simulations of present-day and near-future  
 792 tropospheric ozone, *J. Geophys. Res.*, 111, D08301, doi:10.1029/2005JD006338.  
 793 Thompson, A. M., et al. (2003a), Southern Hemisphere Additional Ozonesondes (SHADOZ)  
 794 1998–2000 tropical ozone climatology: 1. Comparison with total ozone mapping  
 795 spectrometer (TOMS), *J. Geophys. Res.*, 108(D2), 8238, doi:10.1029/2001JD000967.

796 Thompson, A. M., et al. (2003b), Southern Hemisphere Additional Ozonesondes (SHADOZ)  
 797 1998-2000 tropical ozone climatology 2. Tropospheric variability and the zonal wave-  
 798 one, *J. Geophys. Res.*, 108(D2), 8241, doi:10.1029/2002JD002241.  
 799 Thompson, A. M., S. J. Oltmans, D. W. Tarasick, P. Von der Gathen, H. G. J. Smit, J. C. Witte  
 800 (2011), Strategic ozone sounding networks: Review of design and accomplishments,  
 801 *Atmos. Environ.*, 45 (13), 2145-2163, doi:10.1016/j.atmosenv.2010.05.002.  
 802 Thompson, A. M., et al. (2012), Southern Hemisphere Additional Ozonesondes (SHADOZ)  
 803 ozone climatology (2005–2009): Tropospheric and tropical tropopause layer (TTL)  
 804 profiles with comparisons to OMI-based ozone products, *J. Geophys. Res.*, 117, D23301,  
 805 doi:10.1029/2011JD016911.  
 806 Tilmes, S., et al. (2012), Technical Note: Ozone sonde climatology between 1995 and 2011:  
 807 Description, evaluation and applications, *Atmos. Chem. Phys.*, 12, 7475-7497,  
 808 doi:10.5194/acp-12-7475-2012.  
 809 Vesanto, J., J. Himberg, E. Alhoniemi, and J. Parhankangas (2000), SOM Toolbox for Matlab 5,  
 810 report, Helsinki Univ. of Technol., Helsinki, Finland.  
 811 World Meteorological Organization (WMO) (1957), Meteorology – A three-dimensional science:  
 812 Second session of the Commission for Aerology, *WMO Bulletin* vol. 4, no. 4, 134–138.  
 813

814 Table 1: CONUS ozonesonde sites used in this study. Latitude and longitude, altitude amsl, record  
815 length and number of profiles used are shown. Note that the Wallops Island site was moved  
816 slightly (from 4 m elevation) to its current location listed in the table in October 1982.

817

<u>Location</u>	<u>Lat/Lon(°)</u>	<u>Altitude (m)</u>	<u>Length of Record</u>	<u># of Profiles</u>
Boulder, CO	40.0/-105.3	1734	1979-2013	1376
Huntsville, AL	34.7/-86.6	196	1999-2012	686
Trinidad Head, CA	40.8/-124.2	20	1997-2012	868
Wallops Island, VA	37.9/-75.5	13	1970-2013	1600

818

819

820

Table 2: Summaries of SOM node seasonal, meteorological, and O<sub>3</sub> characteristics. Because all sites displayed similar characteristics organized by SOM node, all sites are summarized together. Averages and the range of values from the four sites are presented for tropopause height (m), tropospheric column O<sub>3</sub> (DU), and 6 km O<sub>3</sub> mixing ratio anomaly (ppbv).

<u>Node</u>	<u>Percentage</u> <u>of O<sub>3</sub></u> <u>Profiles</u>	<u>Seasonality</u>	<u>Tropopause</u> <u>Height Anomaly</u> <u>(m)</u>	<u>Tropospheric</u> <u>Column O<sub>3</sub></u> <u>Anomaly (DU)</u>	<u>6 km O<sub>3</sub></u> <u>Anomaly (ppbv)</u>
1:	4 – 8%	Winter-Spring	-1170 (-1500, -710)	-0.6 (-3.3, 1.1)	1.1 (-0.8, 2.2)
2:	4 – 6%	Winter-Spring	-1900 (-2140, -1740)	0.5 (-2.1, 5.1)	7.3 (3.8, 13.2)
3:	2 – 4%	Winter-Spring	-3290 (-4150, -2720)	-3.5 (-7.7, -0.1)	11.9 (6.0, 19.5)
4:	13 – 14%	Most/All Months	30 (-290, 310)	-1.5 (-3.3, -0.3)	-2.2 (-3.0, -0.9)
5:	8 – 13%	All Months	-500 (-960, 160)	1.8 (0.8, 2.7)	-0.2 (-3.8, 3.2)
6:	3 – 5%	Varies by Site	-1380 (-1670, -1020)	6.4 (3.1, 9.8)	10.5 (3.6, 16.2)
7:	22 – 31%	Fall-Winter	770 (600, 880)	-6.2 (-8.5, -3.7)	-7.2 (-8.3, -6.6)
8:	18 – 22%	Spring-Fall	800 (670, 1030)	1.5 (0.5, 1.9)	-0.5 (-2.2, 0.5)
9:	8 – 17%	Summer	240 (-150, 910)	7.8 (5.2, 9.5)	8.8 (7.3, 11.0)

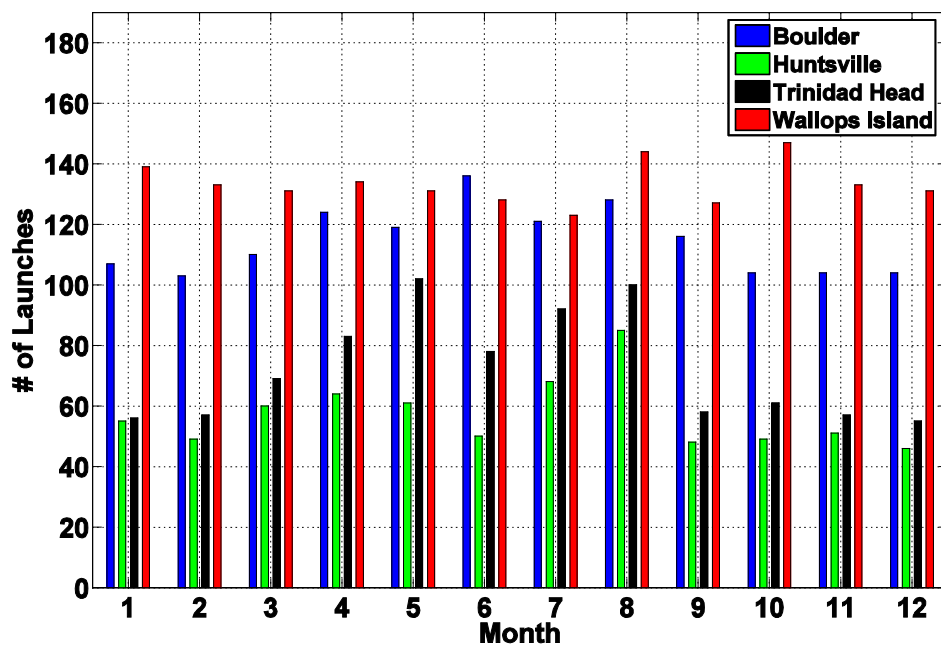


Figure 1: Histogram of number of launches contained in each month for every site.



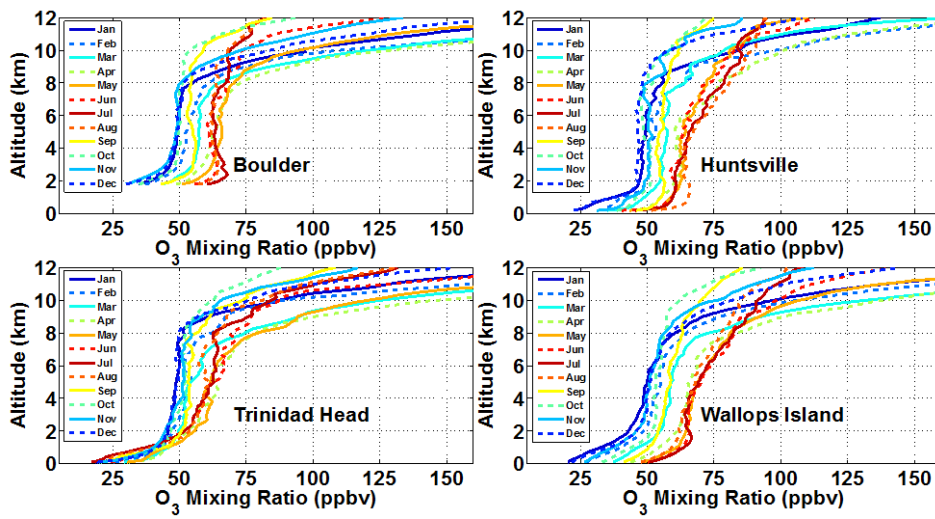


Figure 2: Monthly averaged O<sub>3</sub> mixing ratio profiles for each site from the surface to 12 km amsl.

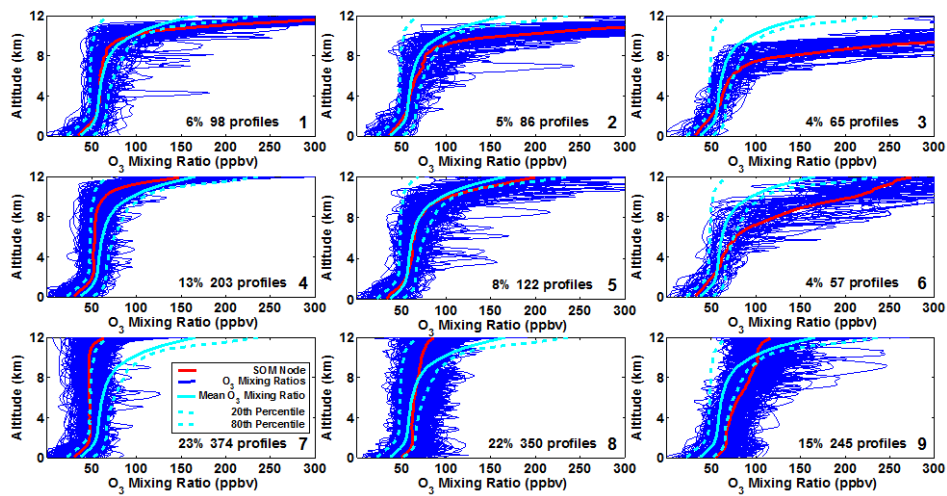


Figure 3: 3x3 SOM output for Wallops Island, VA. SOM nodes are shown in red, with the corresponding individual O<sub>3</sub> mixing ratio profiles in dark blue. For reference, the overall site average O<sub>3</sub> mixing ratio profile, and 20<sup>th</sup> and 80<sup>th</sup> percentile O<sub>3</sub> are shown in cyan.

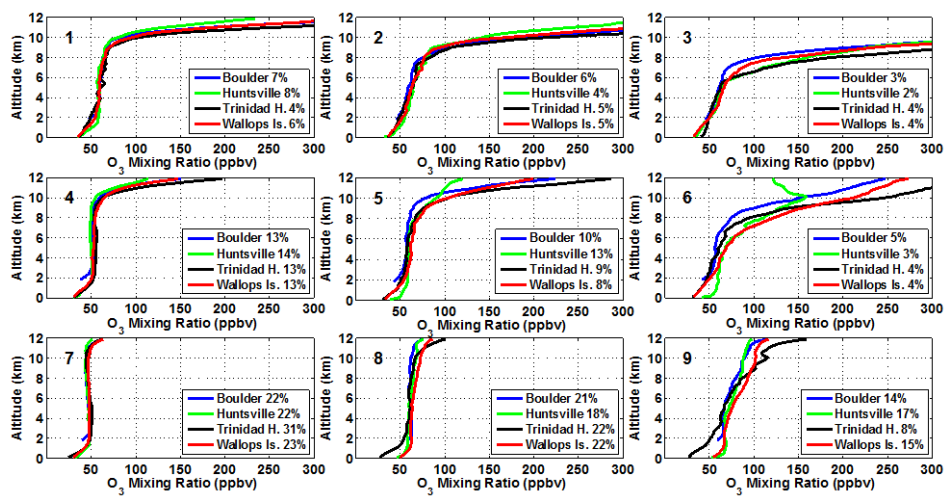


Figure 4: The 3x3 SOM nodes for each of the four CONUS sites shown as O<sub>3</sub> mixing ratio profiles. SOM nodes are labeled from 1 to 9 with the percentage of each site's profiles corresponding to each node shown in the legend.

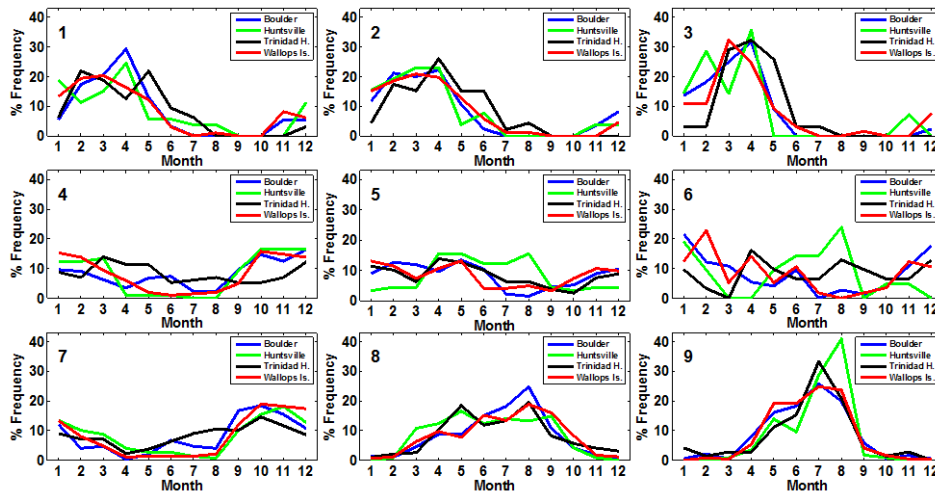


Figure 5: Seasonality of SOM nodes 1 – 9 shown as the relative frequency of month within each SOM node. Each of the nine histograms totals 100% at every site.

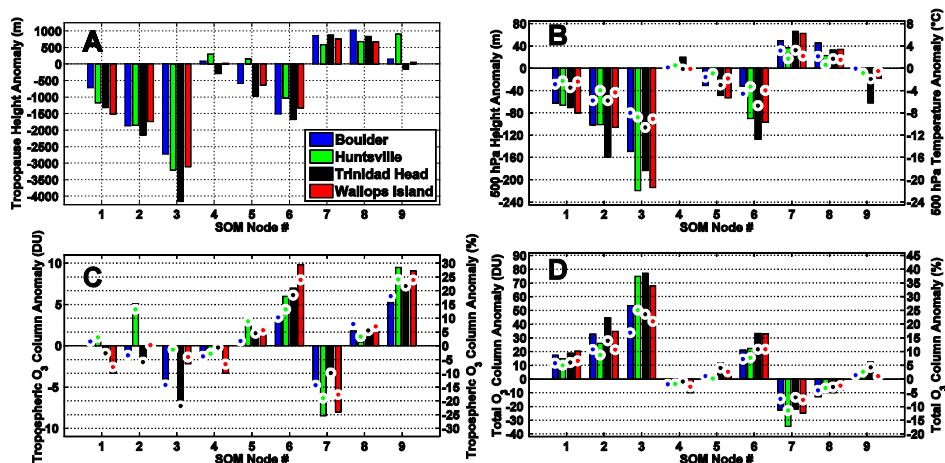


Figure 6: a: Average tropopause height anomaly (in meters) from monthly climatology for each SOM node 1 – 9. b: Average 500 hPa geopotential height anomaly (meters, bars, left axis) and 500 hPa temperature anomaly ( $^{\circ}\text{C}$ , dots, right axis) from monthly climatology for each SOM node 1 – 9. c: Average tropospheric column  $\text{O}_3$  anomaly amount (DU, bars, left axis) and percentage (% , dots, right axis) from monthly climatology for each SOM node 1 – 9. d: Average total column  $\text{O}_3$  anomaly amount (DU, bars, left axis) and percentage (% , dots, right axis) from monthly climatology for each SOM node 1 – 9. Methodology for calculating anomalies is given in Section 3.2.

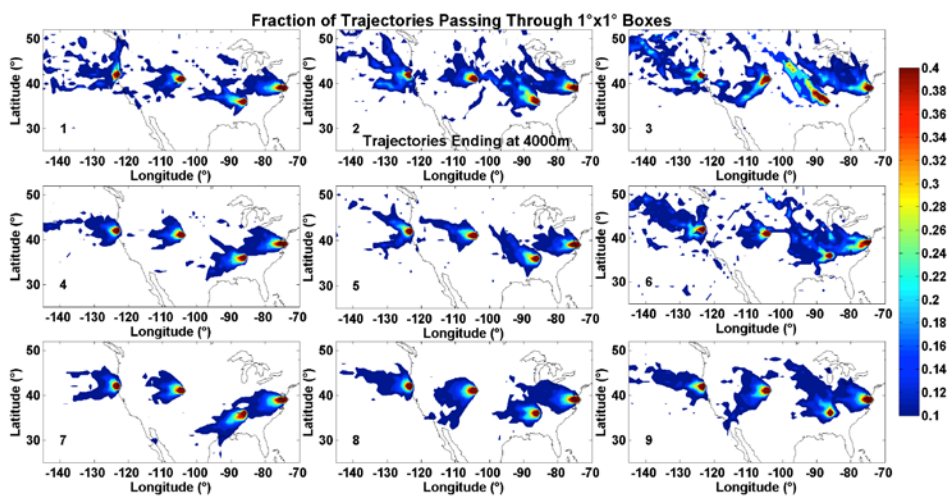


Figure 7: Contoured heat map of HYSPLIT back trajectories terminating at 4000 m at time and location of every  $O_3$  profile. Data are contoured based on fraction of trajectories passing through  $1^\circ \times 1^\circ$  grid boxes. Contours are drawn every 0.02 from 0.10 to 0.40.

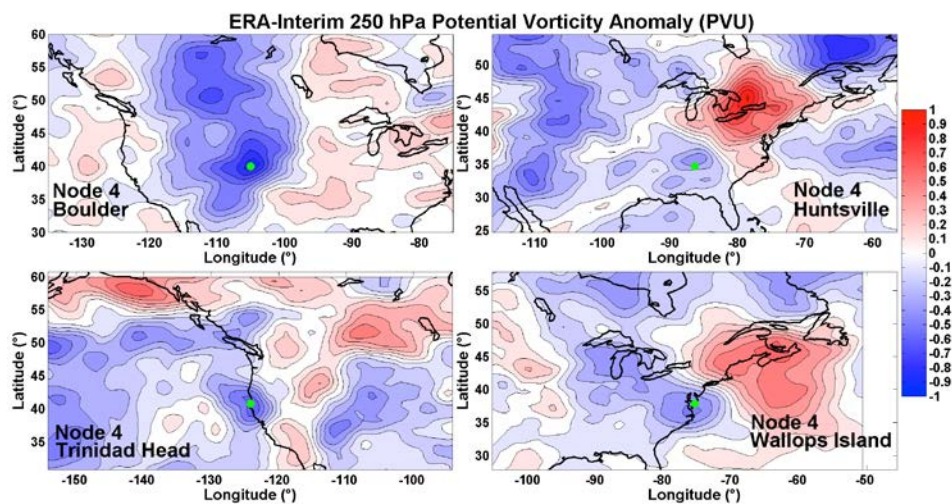


Figure 8: Contoured map of average ERA-Interim 250 hPa PV (PVU) anomalies from monthly climatology for node 4 at each site. Data are contoured every 0.1 PVU from -1 to 1 PVU. Blue colors represent negative anomalies and red colors represent positive anomalies. The green dot represents the site location.

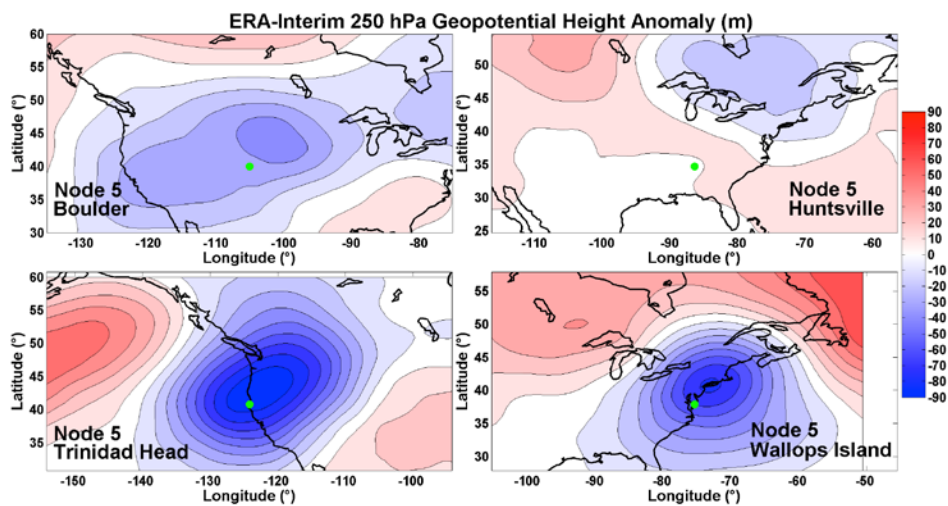


Figure 9: Contoured map of average ERA-Interim 250 hPa geopotential height (m) anomalies from monthly climatology for node 5 at each site. Data are contoured every 10 m from -90 to 90 m. Blue colors represent negative anomalies and red colors represent positive anomalies. The green dot represents the site location.



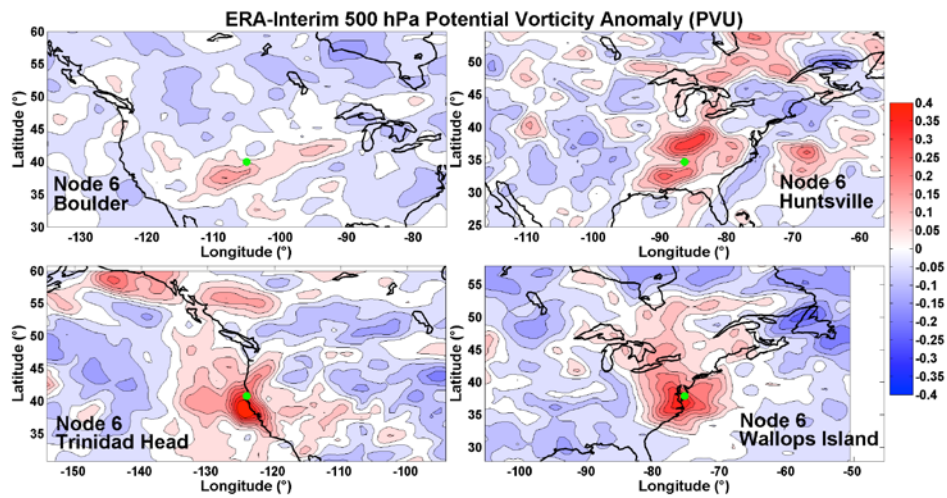


Figure 10: Contoured map of average ERA-Interim 500 hPa PV (PVU) anomalies from monthly climatology for node 6 at each site. Data are contoured every 0.05 PVU from -0.4 to 0.4 PVU. Blue colors represent negative anomalies and red colors represent positive anomalies. The green dot represents the site location.

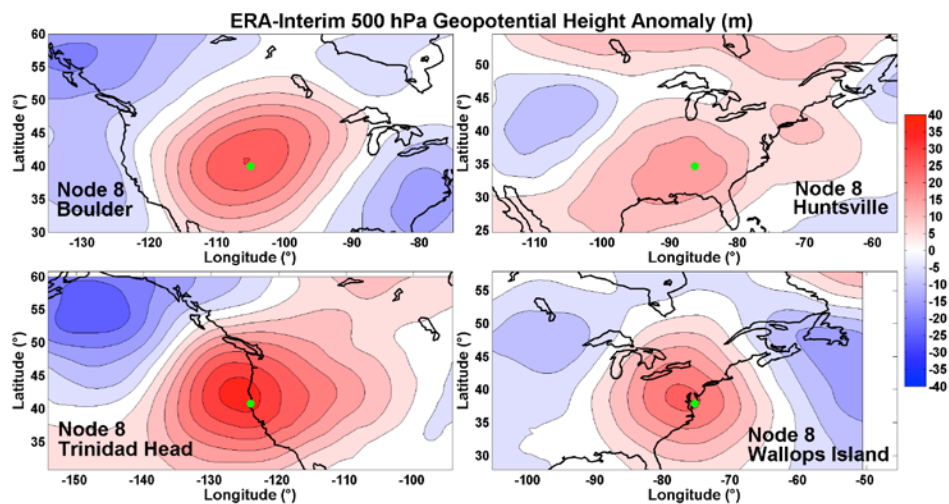
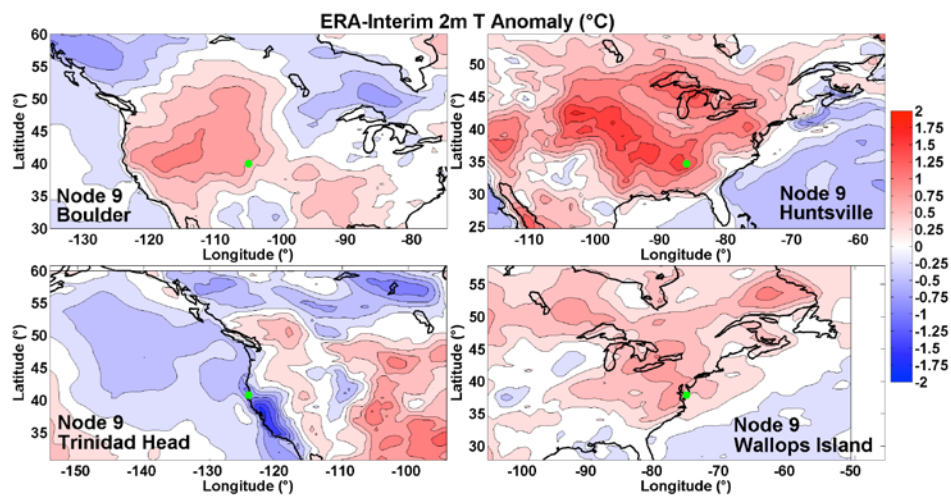


Figure 11: Contoured map of average ERA-Interim 500 hPa geopotential height (m) anomalies from monthly climatology for node 8 at each site. Data are contoured every 5 m from -40 to 40 m. Blue colors represent negative anomalies and red colors represent positive anomalies. The green dot represents the site location.

885



886

887 Figure 12: Contoured map of average ERA-Interim 2 m temperature (°C) anomalies from

888 monthly climatology for node 9 at each site. Data are contoured every 0.25 °C from -2 to 2 °C.

889 Blue colors represent negative anomalies and red colors represent positive anomalies. The green

890 dot represents the site location.

891

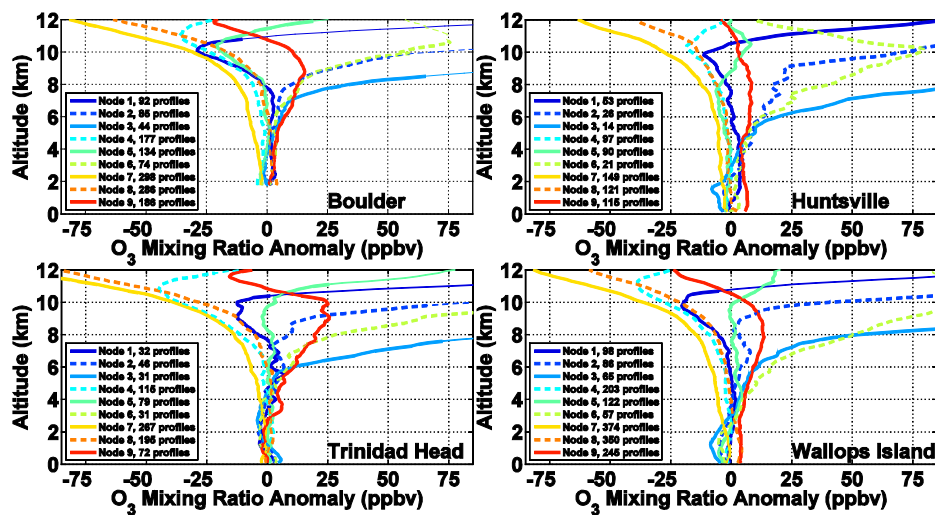


Figure 13: Average O<sub>3</sub> mixing ratio anomaly (ppbv) from monthly climatology with altitude for each 1 – 9 SOM node. For this figure, each O<sub>3</sub> mixing ratio profile was compared with its corresponding monthly averaged O<sub>3</sub> profile. A mean anomaly was then calculated for each node and is shown. For reference, values above the average tropopause for each node are shown as thin lines. The number of profiles in each node is given in the legend.

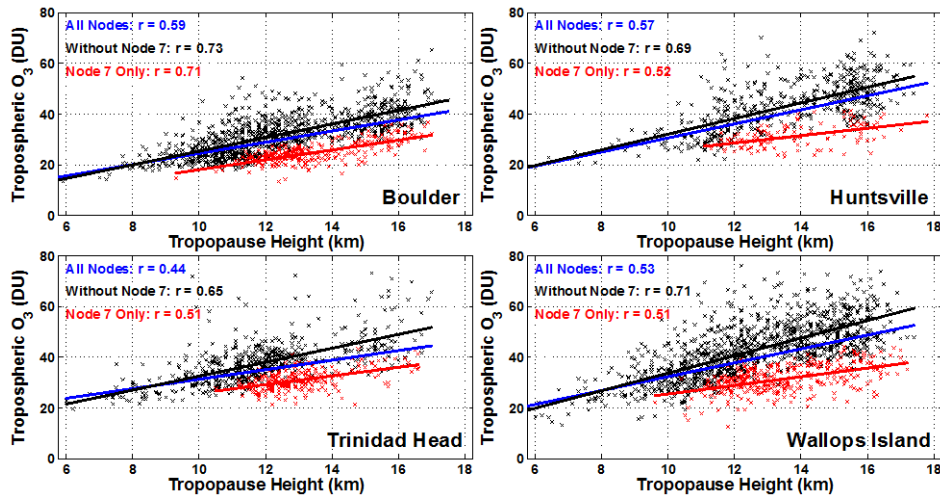


Figure 14: Scatter plots of tropospheric column O<sub>3</sub> (DU) and WMO lapse rate tropopause height (km). Correlation coefficients and least-squares best fit lines are shown for three cases: 1) All profiles/SOM nodes (blue), 2) Excluding node 7 (black), and 3) Only node 7 (red). Individual launches are shown as black crosses with node 7 launches in red.

903 Table A1: Percentages of total profiles corresponding to most and least populous clusters for  
904 given SOM map sizes ( $k$  clusters). Tests were run using SOM with the Epanechnikov  
905 neighborhood function, randomly initialized k-means, and k-means initialized identically to  
906 SOM (via PCA) for 1000 epochs. Note that k-means converges long before 1000 iterations.  
907 Cells marked 0 (1) indicate the cluster contained only one profile.  
908

<u>Site</u>	<u>Map Size (<math>k</math>)</u>	<u>SOM</u>		<u>k-means (rand.)</u>		<u>k-means (PCA)</u>	
		Max%	Min%	Max%	Min%	Max%	Min%
<b>Boulder:</b>	<b>2x2 (4)</b>	44.9	6.3	62.7	3.9	62.7	3.9
	<b>3x3 (9)</b>	21.7	3.2	25.5	2.3	24.6	2.3
	<b>4x4 (16)</b>	12.7	1.5	13.8	0.1	14.5	0.9
<b>Huntsville:</b>	<b>2x2 (4)</b>	46.4	3.6	46.6	3.5	46.6	3.5
	<b>3x3 (9)</b>	21.7	2.0	24.9	1.0	24.5	0.7
	<b>4x4 (16)</b>	15.2	1.0	15.7	0.3	14.6	0 (1)
<b>Trinidad Head:</b>	<b>2x2 (4)</b>	52.1	5.8	62.3	4.4	62.3	4.4
	<b>3x3 (9)</b>	30.8	3.6	31.3	3.1	32.7	3.2
	<b>4x4 (16)</b>	18.0	2.1	11.6	1.6	17.4	0 (1)
<b>Wallops Island:</b>	<b>2x2 (4)</b>	43.5	5.8	46.6	4.7	46.6	4.7
	<b>3x3 (9)</b>	23.4	3.6	25.8	2.0	26.9	1.9
	<b>4x4 (16)</b>	12.4	1.3	14.1	1.4	15.9	1.1

909  
910  
911  
912

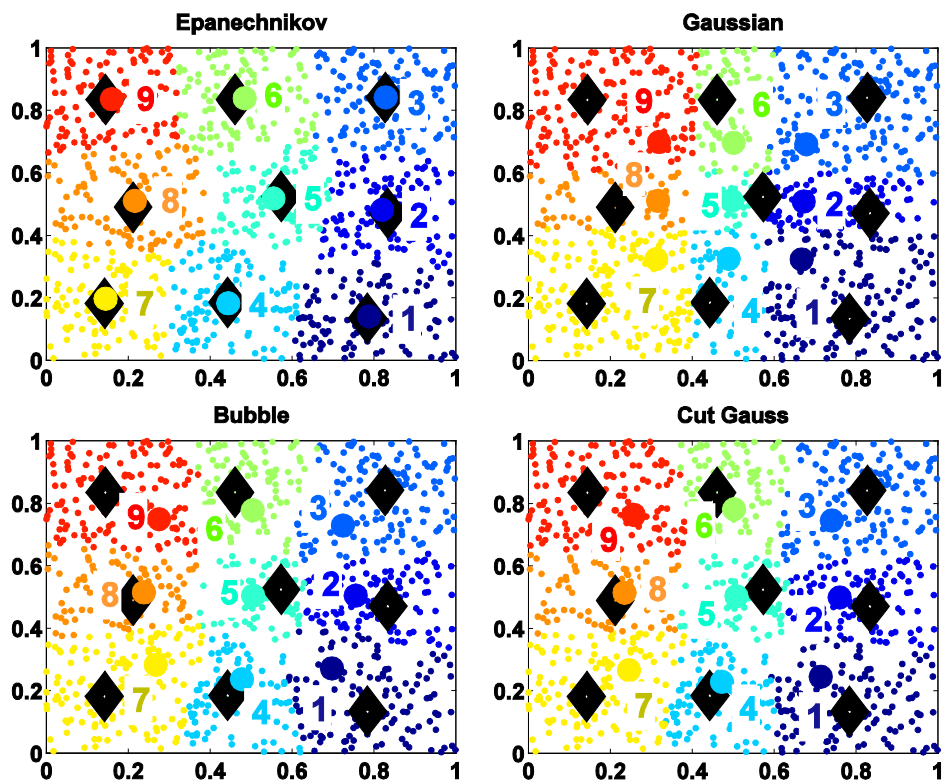


Figure A1: Example of 3x3 SOM nodes (numbered large colored dots) and  $k = 9$  clusters (black diamonds, PCA initialized) of randomly generated data (small colored dots) for four neighborhood functions. Data are colored and labeled according to their respective SOM node in each plot. k-means runs were unchanged and run to convergence. SOM was run for 100 epochs with neighborhood radii decreasing linearly from 3 to 1 over the 100 epochs.

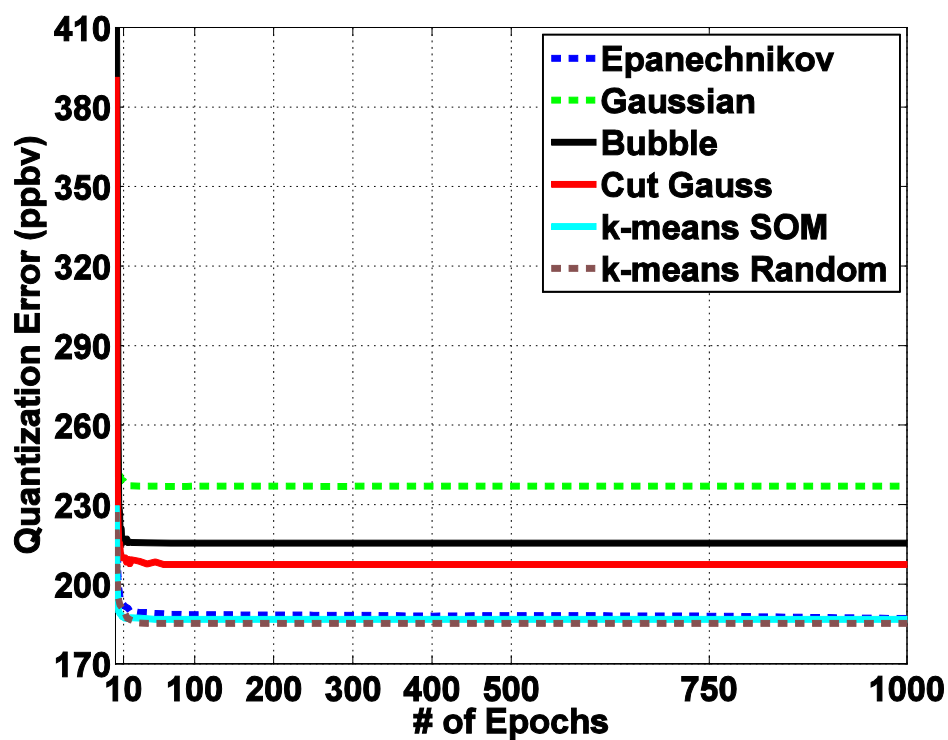


Figure A2: The four-site mean quantization error, defined as the average Euclidean distance between a node/centroid and its member profiles. Data are shown for 3x3 SOM/ $k = 9$  clusters, with increasing epochs. Four SOM neighborhood functions are tested with SOM initialized k-means, and randomly initialized k-means. Note that k-means converges before 100 iterations and is constant thereafter. Lower error indicates a better fit between a node/centroid and its member data.



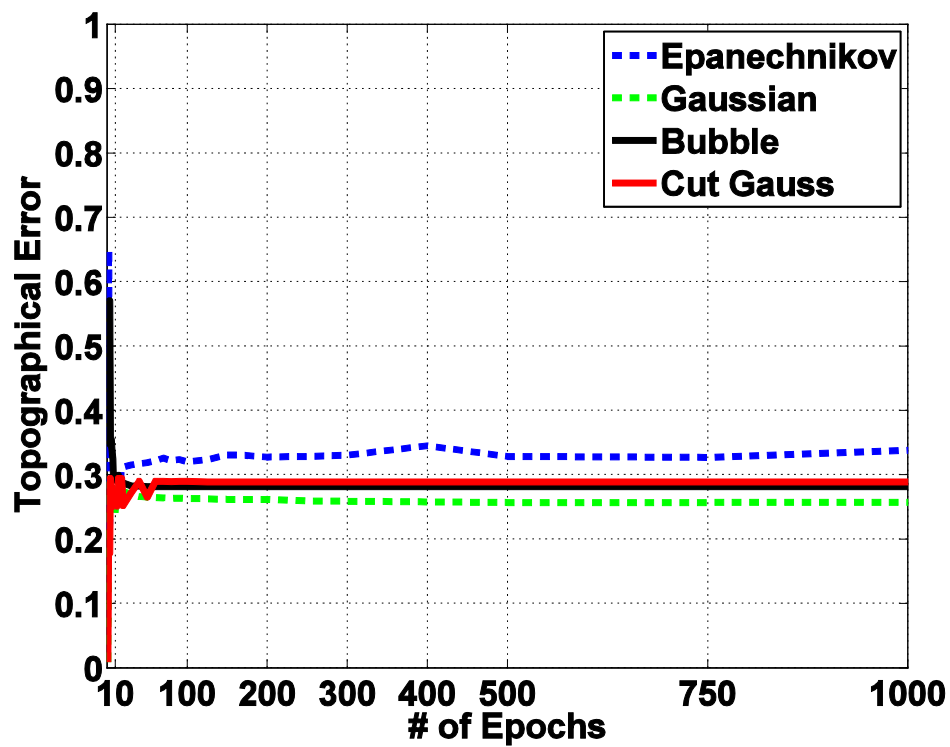


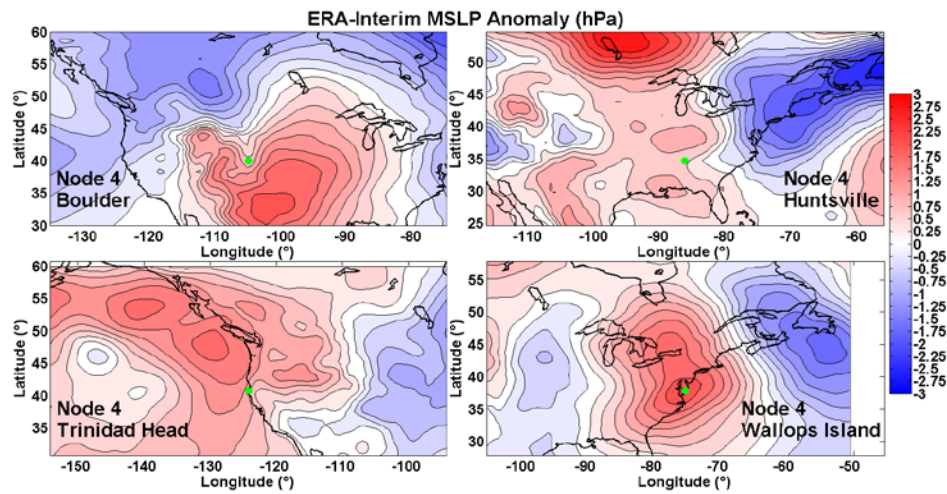
Figure A3: The four-site mean topographical error, defined as the fraction of profiles whose second closest node is *not* adjacent to its BMU in the map, for 3x3 SOM,  $k = 9$  clusters, with number of epochs for four SOM neighborhood functions. Data are from the same output as in Figure A2. Lower percentages typically indicate better-ordered neighboring nodes.

934 Table S1: Percentage of 3x3 SOM node profiles that are members of 2x2 SOM node clusters for  
935 each CONUS site. For example, 86% of the profiles in the Boulder 3x3 SOM node 6 are members  
936 of the 2x2 SOM node 1. Note that every node 7 profile, regardless of site, is a member of the 2x2  
937 SOM node 3.

	Node	Node	Node	Node	Node	Node	Node	Node	Node
	1 (%)	2 (%)	3 (%)	4 (%)	5 (%)	6 (%)	7 (%)	8 (%)	9 (%)
<b>Boulder</b>									
2x2 Node 1	98	49	0	0	11	86	0	0	0
2x2 Node 2	0	51	100	0	0	0	0	0	0
2x2 Node 3	0	0	0	36	0	0	100	89	1
2x2 Node 4	2	0	0	64	89	14	0	11	99
<b>Huntsville</b>									
2x2 Node 1	96	65	0	1	2	5	0	0	0
2x2 Node 2	0	35	100	0	0	10	0	0	0
2x2 Node 3	0	0	0	97	2	0	100	60	0
2x2 Node 4	4	0	0	2	96	86	0	40	100
<b>Trinidad</b>									
<b>Head</b>									
2x2 Node 1	100	59	0	0	16	97	0	0	0
2x2 Node 2	0	41	100	0	0	0	0	0	0
2x2 Node 3	0	0	0	16	0	0	100	85	1
2x2 Node 4	0	0	0	84	84	3	0	15	99
<b>Wallops</b>									
<b>Island</b>									
2x2 Node 1	94	73	0	0	12	89	0	0	0
2x2 Node 2	0	27	100	0	0	7	0	0	0
2x2 Node 3	0	0	0	67	0	0	100	53	0
2x2 Node 4	6	0	0	33	88	4	0	47	100

938  
939

940



941

942

943

944

945

946

Figure S1: Contoured map of average ERA-Interim MSLP (hPa) anomalies from monthly climatology for node 4 at each site. Data are contoured every 0.25 hPa from -3 to 3 hPa. Blue colors represent negative anomalies and red colors represent positive anomalies. The green dot represents the site location.

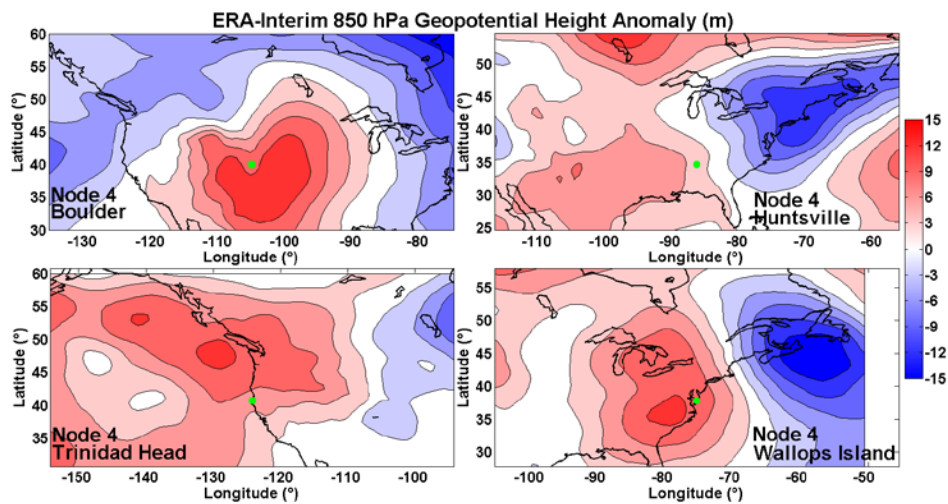


Figure S2: Contoured map of average ERA-Interim 850 hPa geopotential height (m) anomalies from monthly climatology for node 4 at each site. Data are contoured every 3 m from -15 to 15 m. Blue colors represent negative anomalies and red colors represent positive anomalies. The green dot represents the site location.